The Dissertation Committee for Ali Jalali
certifies that this is the approved version of the following dissertation:

Dirty Statistical Models

Committee:

Sujay Sanghavi, Supervisor

Constantine Caramanis

Inderjit Dhillon

Joydeep Ghosh

Pradeep Ravikumar

Dirty Statistical Models

by

Ali Jalali, B.S.; M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2012

Dedicated to Soltän.

Acknowledgments

I wish to thank all professors, students and staff of The Wireless Networking and Communication Networking Group at UT Austin for their kind helps throughout my PhD. In particular, I would like to thank my adviser, Prof. Sujay Sanghavi, whose constant support was always encouraging me to learn more and without his dedication, this thesis was impossible. I would also like to express my gratitude to Prof. Pradeep Ravikumar for insightful discussions that helped me find my research path. My special thanks to my parents and little sister for their unconditional love and support during my hardest times. I am running out of words when it comes to thanking my friends, my family by choice, for being always with me in difficulties.

Dirty Statistical Models

Publication No. _____

Ali Jalali, Ph.D.
The University of Texas at Austin, 2012

Supervisor: Sujay Sanghavi

In fields across science and engineering, we are increasingly faced with problems where the number of variables or features we need to estimate is much larger than the number of observations. Under such high-dimensional scaling, for any hope of statistically consistent estimation, it becomes vital to leverage any potential structure in the problem such as sparsity, low-rank structure or block sparsity. However, data may deviate significantly from any one such statistical model. The motivation of this thesis is: can we simultaneously leverage more than one such statistical structural model, to obtain consistency in a larger number of problems, and with fewer samples, than can be obtained by single models? Our approach involves combining via simple linear superposition, a technique we term *dirty models*. The idea is very simple: while any one structure might not capture the data, a superposition of structural classes might. Dirty models thus searches for a parameter that can be *decomposed* into a number of simpler structures such as (a) sparse plus block-sparse, (b) sparse plus low-rank and (c) low-rank plus block-sparse. In this thesis, we propose dirty model based algorithms for different problems such as multi-task learning, graph clustering and time-series analysis with latent factors. We analyze these algorithms in terms of the number of observations we need to estimate the variables. These algorithms are based on convex optimization and sometimes they are relatively slow. We provide a class of low-complexity greedy algorithms that not only can solve these optimizations faster, but also guarantee the solution. Other than theoretical results, in each case, we provide experimental results to illustrate the power of dirty models.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In many applications such as pattern recognition, machine vision, bioinformatics, data mining, financial engineering, etc, we face parameter estimation problems where the number of observations is much less than the number of the dimension of the parameter. In such a high-dimensional regime, there is no hope for consistent estimation unless we restrict the parameter to have a certain structure model. In other words, the true parameter lives in a high-dimensional space, but the observed data has much lower intrinsic dimensions. The hope is that few observations suffice to recover the parameter in the low-dimensional sub-space. Common, and recently popular, structure models are

- **Sparsity**: This model suggests that most entries of the unknown parameter are zero and there are only few non-zero elements. This structure has been studied in many areas such as compressed sensing [13] and LASSO [131].
- **Block Sparse**: This model partitions the entries of the unknown parameter into a number of subsets (blocks) and suggests that within each block, either all entries are zero or most of them, maybe all, are non-zero [26, 111].
- **Low-Rank**: This model suggests that the parameter matrix has a low-dimensional row/column space and hence has much lower rank compare to its size [102, 116].
- **Sparse Markov Random Field**: This model suggests that the graphical model associated with a set of random variables is sparse [112, 113]; that is the joint probability distribution of a set of random variables can be factorized such that each factor is the joint probablity distribution of a subset of those random variables.

In each of these models, if we knew apriori the exact instance of the overall dimensional structure, inference is easy. The challenge however is finding the correct instance of the structure from a very large number of possibilities. For example, in sparse linear regression, if we know the location of non-zero entries, then we can set the other entries to zero and estimate only non-zero entries, say by solving a convex optimization problem. Similarly, in block sparse model, if we know which blocks are non-zero, or in low-rank model, if we know the singular vectors of the matrix, then we can estimate non-zero entries with very few observations. Hence, recovering the structure of a model is the critical task in high-dimensional setting. Once the structure of the model is recovered, estimation is a fairly easy job.

There is a hidden assumption behind choosing a certain model for the unknown parameter. In fact, we assume that the underlying true data structure falls into one of these models completely, i.e., the data is *clean*. Notice that when we use the term clean data, we do not necessarily mean that there is no niose. For example, if x is a vector, then $Q = xx^T + w$, with additive noise w, can be considered to have low-rank model (which is a clean model). However, in many real applications, we face more complex structures, where the underlying data structure does not fit any model solely. Instead, the data structure model can be expressed as the superposition of a number of simpler models, i.e., the data is *dirty*. For example, we can cosider the aformentioned matrix Q to be a superposition of low-rank and sparse models, assuming that the noise has only few non-zero elements. We call such a superposition of simple models a *dirty model*.

Our main idea in this thesis is to analyze dirty models for dirty data and try to recover the structure of each of these models. We have analyzed recovery of different dirty models via convex optimization and greedy algorithms in terms of their sample complexity, structure recovery guarantees and comparison to other clean models. As discussed before, once the structure is revealed, the estimation is easy because the unknown parameter will be restricted to a low-dimensional subspace.

1.1 Structure Recovery Techniques

In this section, we discuss two main structure recovery techniques and their properties. Instances of each of these techniques as well as relevant applications are presented in the consequent chapters of this thesis.

1.1.1 Convex Optimization

The use of convex optimization for structure recovery is very broad and popular. At a high level, given a set of observation D, we would like to estimate a target parameter θ^*. There are two ingredients for a consistent convex optimization based structure recovery algorithm as follows:

- **Loss Function:** We need a convex loss function $\mathcal{L}(\theta; D)$ that given our observation, it assigns a *penalty* to each parameter θ. This loss function must have θ^* as the asymptotically optimal point, i.e., $\mathbb{E}_D\left[\nabla \mathcal{L}(\theta^*; D)\right] = 0$. Moreover, the more curvature the loss function around θ^* has, the easier the estimation will be.

- **Regularizer:** In a high-dimensional setting, to get a consistent estimation, we assume that $\theta^* \in \mathcal{C}$ for some structure set $\mathcal{C}$. Examples of such a set are the set of 1-sparse vectors or the set of 1-rank matrices. We then construct a function $R_{\mathcal{C}}(\theta)$ to be the function that encourages to pick elements from the set $\mathcal{C}$ as opposed to an element from outside, i.e., $R_{\mathcal{C}}(\theta^*) \leq R_{\mathcal{C}}(\theta)$ for any $\theta \notin \mathcal{C}$. This function is often referred to as *regularizer*. Since the set $\mathcal{C}$ and consequently the regularizer $R_{\mathcal{C}}$ are often not-convex, we pick the convex envelope $\bar{R}_{\mathcal{C}}$ to ensure tractability.

Using these two ingredients, we solve the following convex program to get an estimate of θ^*.

$$\widehat{\theta} = \arg\min_{\theta} \ \mathcal{L}(\theta; D) + \lambda \bar{R}_{\mathcal{C}}(\theta).$$

The parameter λ can be chosen via techniques like cross-validation or bootstrapping. This class of algorithms are known to have two main issues:

- **Robustness:** A single observation can arbitrarily change $\widehat{\theta}$ and hence, the method is not robust with respect to outliers.
- **Flexibility:** If θ^* represents a multi-modal data, then $\underline{\theta}$ potentially represents an *average* mode which is not a representative of neither modes of the data.

Using our dirty model, we can fight against both issues. Imagine $\theta^* = \theta_1^* + \theta_2^*$ where θ_1^* and θ_2^* come from two structures $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. To get robustness, we choose $\mathcal{C}_1$ to be the structure of our data and $\mathcal{C}_2$ to be the structure of the outliers. Similarly, to get flexibility, we choose $\mathcal{C}_1$ and $\mathcal{C}_2$ to represent different modes of the data. Subsequently, we solve the following convex optimization problem:

$$(\widehat{\theta}_1, \widehat{\theta}_2) = \arg\min_{\theta_1,\theta_2} \mathcal{L}(\theta_1 + \theta_2; D) + \lambda_1 \bar{R}_{\mathcal{C}_1}(\theta_1) + \lambda_2 \bar{R}_{\mathcal{C}_2}(\theta_2).$$

In the robust case, we output $\widehat{\theta} = \widehat{\theta}_1$ and in the flexible case, we output $\widehat{\theta} = \widehat{\theta}_1 + \widehat{\theta}_2$. In this thesis, we investigate some instances of this problem for both robustness (Chapters 3,4,5) and flexibility (Chapter 2) and show the advantages of this dirty model based algorithm over simple model counterparts.

One of the difficulties with robust dirty model analysis is to characterize different structures so that without any ambiguity they can be distinguished. Recall the example of low-rank and sparse models superposition and notice that a low-rank matrix can be also sparse or conversely, a sparse matrix can be low-rank. Thus, we need to not only reduce the size of the problem by imposing the structure, but also further restricting each structure to be consistently incoherent from each other to get robustness. Without this separation, the problem is not well-defined and characterization of the models is impossible. Motivated by this, we normally impose some conditions that uniquely partitions the space of parameters so that each non-zero element of the space belongs to only one structure.

1.1.2 Greedy Algorithms

Considering the original non-convex regularizer $R_{\mathcal{C}}(\cdot)$, the *quality* of the convex optimization method depends on the tightness of the convex relaxation $\bar{R}_{\mathcal{C}}(\cdot)$. Often to ensure tightness, convex optimization problems require strong assumptions such as irrepresentable condition that imposes strong restrictions on the loss function. Besides, convex optimization methods, although polynomial, are still computationally expensive. In this thesis, we show that greedy algorithms are good candidates to both reduce the computational complexity and relax strong assumptions while maintaining the same guarantees as convex optimization.

The main idea of greedy algorithms is basis pursuit and in particular orthogonal matching pursuit [135, 154] – that is iteratively find the next best "coordinate" in $\mathcal{C}$ and add it to the estimated structure. This simple forward greedy algorithm has been shown to perform as good as the convex optimization. However, it still requires the same strong assumptions. Like many forward algorithms, the beginning steps are far more important than the others and if the first steps are mistaken, then the algorithm performs poorly.

We introduce a forward-backward greedy algorithm to solve sparse + block-sparse dirty model based on the algorithms introduced in [68, 152]. We show this algorithm not only lowers the sample complexity, but also requires only mild assumption of restricted strong convexity [100]. This algorithm greedily picks the best coordinates in the forward step and then removes "bad" coordinates in the backward step. We show both theoretically and empirically outperforms the corresponding convex optimization method.

1.2 Studied Dirty Models

We studied different instances of two major classes of dirty models introduced in the next two sections. These dirty models include sparse+block-sparse and sparse+low-rank models.

1.2.1 Sparse + Block Sparse

Suppose we have multiple linear regression problems, i.e., multi-task learning problem [26]. Here, multiple tasks share some common structure such as sparsity, and estimating these tasks jointly by leveraging this common structure could be more statistically efficient. We have $r > 1$ response variables (tasks), and a common set of p features or covariates. The setting we focus on is where the response variables have *simultaneously sparse* structure: the index set of relevant features for each task is sparse; and there is a large overlap of these relevant features across the different regression problems. Such "simultaneous sparsity" arises in a variety of contexts [133]; indeed, most applications of sparse signal recovery in contexts ranging from graphical model learning, kernel learning, and function estimation have natural extensions to the simultaneous-sparse setting [7, 109, 111].

It is useful to represent the multiple regression parameters via a matrix, where each column corresponds to a task, and each row to a feature. Having simultaneous sparse structure then corresponds to the matrix being largely "block-sparse" – where each row is either all zero or mostly non-zero, and the number of non-zero rows is small. A lot of recent research in this setting has focused on ℓ_1/ℓ_q norm regularizations, for $q > 1$, that encourage the parameter matrix to have such block-sparse structure. Particular examples include results using the ℓ_1/ℓ_∞ norm [101, 136, 151], and the ℓ_1/ℓ_2 norm [89, 104].

Our method searches for a parameter matrix that can be *decomposed* into a row-sparse matrix (corresponding to the overlapping or shared features) and an elementwise sparse matrix (corresponding to the non-shared features). As we show both theoretically and empirically, with this simple fix we are able to leverage any extent of shared features, while allowing disparities in support and values of the parameters, so that we are *always* better than both the Lasso or block-sparse regularizers (at times remarkably so). In Chapter 2, we provide a convex optimization based algorithm and in Chapter 6, we provide a forward-backward greedy algorithm for the same problem.

1.2.2 Sparse + Low-Rank

In many applications such as principal component analysis, often times we want to recover a low-rank matrix and a sparse matrix from their sum. Additive large-magnitude noise potentially can change the rank of the matrix by much. In such a heavily noisy regime, PCA based approaches has poor performance. Thus, recovering the underlying low-rank structure in presence of the noise is a challenging job. We model this problem using dirty models and cast it as a convex optimization problem. We consider graph clustering and time-series analysis as applications of using this dirty model.

1.2.2.1 Graph Clustering

The problem we look at is: given an unweighted graph, partition the nodes, i.e., cluster, so as to minimize the sum of (a) number of edges between endpoints that are in different partitions, and (b) the number of missing edges between endpoints in the same partition. This is one particular non statistical application of sparse and low rank matrix separation. Given a clustering, we call edges of type (a) or (b) "disagreements"; we are thus interested in optimal clusterings – those that minimize the number of disagreements. This formulation has the advantage of not requiring an external parametric input of how many clusters there should be in the final solution; this is fully determined by the data at hand.

This problem, as formulated above, is exactly the same as the problem of correlation clustering, first proposed by Bansal, Blum and Chawla [10]. Specifically, edges that [10] labels as "+" are those that are "present in the graph" for us, and those labelled as "-" in [10] are those that are "missing". They consider the problem with +/- labels on the complete graph. [10] showed that finding the exact optimum is NP-complete; they then proceed to provide a constant-factor approximation algorithm for minimizing the number of disagreements.

We take an alternative approach to the problem; instead of looking for an approximation that holds for all problem inputs, we provide algorithms that either (a) yields the optimum (i.e. disagreement minimizing) clustering, or (b) generates a FAILURE flag, yielding no clustering. Our algorithm is

based on using recently developed matrix uncertainty principles [22, 31] - that (certain classes of) matrices cannot be simultaneously sparse and low-rank. For these matrices, we can *exactly* recover the sparse matrix ($\mathbf{B}$) and low-rank matrix ($\mathbf{K}$) given only their sum ($\mathbf{B}$+$\mathbf{K}$), using convex optimization. In the graph context, the low-rank matrix corresponds to the cliques that would be an ideal input corresponding to the optimal clustering, the composite matrix represents the given data, and the sparse matrix represents the disagreements. Existing results on sparse and low-rank matrix decompositions provide weak guarantees for the graph clustering case; we consider two types of relaxation based on nuclear norm in Chapter 3 and max-norm in Chapter 4 and provide much stronger guarantees.

1.2.2.2 Time-Series Analysis

Suppose we have a non-stationary random vector whose mean and variance evolves over the time and we observe some entries of this vector for a finite time. The evolution of the mean and variance of each entry depends on both observed and unobserved (latent) entries. The problem we would like to investigate is that whether or not it is possible to learn the evolution process of the mean and variance of the observed entries. We consider linear stochastic first-order evolution model $\dot{x}(t) = Ax(t) + \dot{B}(t)$, where x is the random vector, A is the evolution coefficients matrix and $B(t)$ is the standard Brownian motion noise, and learn the matrix A; the problem that is investigated for the case when there is no latent variable in [15].

We formulate this problem as a sparse plus low-rank dirty model of $A = S+L$, where, S captures the effect of observed variables on each other and L captures the effect of latent variables on observed entries. Here, we assume that few observed entries affect each observed entry and also each observed entry affects only few observed entries, and hence, S is sparse. Moreover, we show that if the number of latent variables is less than observed entries, then the matrix L is low-rank. Unlike most cases of studies in sparse and low-rank decomposition, here, the focus is to learn the structure of the sparse matrix S rather than the low-rank matrix L.

There are similarities between this problem and learning Gaussian graphical models with latent variables [28]. However, at least in [28] the focus has

been to recover the number of latent variables, which is the rank of L, and more importantly, in graphical model learning, the assumption is that the samples are independent and law of large numbers can be applied easily. We need to provide much more subtle analysis since because of the time evolution, our samples are highly correlated and hence, the usual concentration results (of independent variables) cannot be applied. In Chapter 5, we investigate this problem and provide guarantees of success for our method. Further, we apply our method to stock market prediction problem and observe significant improvements.

Chapter 2

A Dirty Model for Multiple Sparse Regression

Sparse linear regression – finding an unknown vector from linear measurements – is now known to be possible with fewer samples than variables, via methods like the LASSO. We consider the multiple sparse linear regression problem, where several related vectors – with *partially shared* support sets – have to be recovered. A natural question in this setting is whether one can use the sharing to further decrease the overall number of samples required. A line of recent research has studied the use of ℓ_1/ℓ_q norm block-regularizations with $q > 1$ for such problems; however these could actually perform *worse* in sample complexity – vis a vis solving each problem separately ignoring sharing – depending on the level of sharing.

We present a new method for multiple sparse linear regression that can leverage support and parameter overlap when it exists, but not pay a penalty when it does not. a very simple idea: we decompose the parameters into two components and *regularize these differently*. We show both theoretically and empirically, our method strictly and noticeably outperforms both ℓ_1 or ℓ_1/ℓ_q methods, over the entire range of possible overlaps (except at boundary cases, where we match the best method). We also provide theoretical guarantees that the method performs well under high-dimensional scaling.

2.1 Introduction: Motivation and Setup

High-dimensional scaling. In fields across science and engineering, we are increasingly faced with problems where the number of variables or features p is larger than the number of observations n. Under such high-dimensional scaling, for any hope of statistically consistent estimation, it becomes vital to leverage any potential structure in the problem such as sparsity (e.g. in

compressed sensing [13] and LASSO [131]), low-rank structure [102, 116], or sparse graphical model structure [111]. It is in such high-dimensional contexts in particular that multi-task learning [26] could be most useful. Here, multiple tasks share some common structure such as sparsity, and estimating these tasks jointly by leveraging this common structure could be more statistically efficient.

Block-sparse Multiple Regression. A common multiple task learning setting, and which is the focus of this chapter, is that of multiple regression, where we have $r > 1$ response variables, and a common set of p features or covariates. The r tasks could share certain aspects of their underlying distributions, such as common variance, but the setting we focus on in this chapter is where the response variables have *simultaneously sparse* structure: the index set of relevant features for each task is sparse; and there is a large overlap of these relevant features across the different regression problems. Such "simultaneous sparsity" arises in a variety of contexts [133]; indeed, most applications of sparse signal recovery in contexts ranging from graphical model learning, kernel learning, and function estimation have natural extensions to the simultaneous-sparse setting [7, 109, 111].

It is useful to represent the multiple regression parameters via a matrix, where each column corresponds to a task, and each row to a feature. Having simultaneous sparse structure then corresponds to the matrix being largely "block-sparse" – where each row is either all zero or mostly non-zero, and the number of non-zero rows is small. A lot of recent research in this setting has focused on ℓ_1/ℓ_q norm regularizations, for $q > 1$, that encourage the parameter matrix to have such block-sparse structure. Particular examples include results using the ℓ_1/ℓ_∞ norm [101, 136, 151], and the ℓ_1/ℓ_2 norm [89, 104].

Our Model. Block-regularization is "heavy-handed" in two ways. By strictly encouraging shared-sparsity, it assumes that all relevant features are shared, and hence suffers under settings, arguably more realistic, where each task depends on features specific to itself in addition to the ones that are common. The second concern with such block-sparse regularizers is that the ℓ_1/ℓ_q norms can be shown to encourage the entries in the non-sparse rows taking nearly identical *values*. Thus we are far away from the original goal of multitask learning: not only do the set of relevant features have to be exactly the same,

but their values have to as well. Indeed recent research into such regularized methods [101, 104] caution against the use of block-regularization in regimes where the supports and values of the parameters for each task can vary widely. Since the true parameter values are unknown, that would be a worrisome caveat.

We thus ask the question: can we learn multiple regression models by leveraging whatever overlap of features there exist, and without requiring the parameter values to be near identical? Indeed this is an instance of a more general question on whether we can estimate statistical models where the data may not fall cleanly into any one structural bracket (sparse, block-sparse and so on). With the explosion of complex and *dirty* high-dimensional data in modern settings, it is vital to investigate estimation of corresponding *dirty* models, which might require new approaches to biased high-dimensional estimation. In this chapter we take a first step, focusing on such dirty models for a specific problem: simultaneously sparse multiple regression.

Our approach uses a simple idea: while any one structure might not capture the data, a superposition of structural classes might. Our method thus searches for a parameter matrix that can be *decomposed* into a row-sparse matrix (corresponding to the overlapping or shared features) and an elementwise sparse matrix (corresponding to the non-shared features). As we show both theoretically and empirically, with this simple fix we are able to leverage any extent of shared features, while allowing disparities in support and values of the parameters, so that we are *always* better than both the Lasso or block-sparse regularizers (at times remarkably so).

Notation: For any matrix M, we denote its j^{th} row as m_j, and its k-th column as $m^{(k)}$. The set of all non-zero rows (i.e. all rows with at least one non-zero element) is denoted by $\text{RowSupp}(M)$ and its support by $\text{Supp}(M)$. Also, for any matrix M, let $\|M\|_{1,1} := \sum_{j,k} |m_j^{(k)}|$, i.e. the sums of absolute values of the elements, and $\|M\|_{1,\infty} := \sum_j \|m_j\|_\infty$ where, $\|m_j\|_\infty := \max_k |m_j^{(k)}|$.

2.2 Problem Set-up and Our Method

Multiple regression. We consider the following standard multiple linear regression model:

$$y^{(k)} = X^{(k)}\bar{\theta}^{(k)} + w^{(k)}, \quad k = 1, \ldots, r,$$

where, $y^{(k)} \in \mathbb{R}^n$ is the response for the k-th task, regressed on the design matrix $X^{(k)} \in \mathbb{R}^{n \times p}$ (possibly different across tasks), while $w^{(k)} \in \mathbb{R}^n$ is the noise vector. We assume each $w^{(k)}$ is drawn independently from $\mathcal{N}(0, \sigma^2)$. The total number of tasks or target variables is r, the number of features is p, while the number of samples we have for each task is n. For notational convenience, we collate these quantities into matrices $Y \in \mathbb{R}^{n \times r}$ for the responses, $\bar{\Theta} \in \mathbb{R}^{p \times r}$ for the regression parameters and $W \in \mathbb{R}^{n \times r}$ for the noise.

Our Model. In this chapter we are interested in estimating the true parameter $\bar{\Theta}$ from data $\{y^{(k)}, X^{(k)}\}$ by leveraging any (unknown) extent of simultaneous-sparsity. In particular, certain rows of $\bar{\Theta}$ would have many non-zero entries, corresponding to features shared by several tasks ("shared" rows), while certain rows would be elementwise sparse, corresponding to those features which are relevant for some tasks but not all ("non-shared rows"), while certain rows would have all zero entries, corresponding to those features that are not relevant to any task. We are interested in estimators $\widehat{\Theta}$ that automatically adapt to different levels of sharedness, and yet enjoy the following guarantees:

Support recovery: We say an estimator $\widehat{\Theta}$ successfully recovers the true signed support if $\text{sign}(\text{Supp}(\widehat{\Theta})) = \text{sign}(\text{Supp}(\bar{\Theta}))$. We are interested in deriving sufficient conditions under which the estimator succeed. We note that this is stronger than merely recovering the row-support of $\bar{\Theta}$, which is union of its supports for the different tasks. In particular, denoting $\mathcal{U}_k$ for the support of the k-th column of $\bar{\Theta}$, and $\mathcal{U} = \bigcup_k \mathcal{U}_k$.

Error bounds: We are also interested in providing bounds on the elementwise ℓ_∞ norm error of the estimator $\widehat{\Theta}$,

$$\|\widehat{\Theta} - \bar{\Theta}\|_\infty = \max_{j=1,\ldots,p} \max_{k=1,\ldots,r} \left| \widehat{\Theta}_j^{(k)} - \bar{\Theta}_j^{(k)} \right|.$$

Algorithm 1 Complex Block Sparse

Solve the following convex optimization problem:

$$(\widehat{S}, \widehat{B}) \in \arg\min_{S,B} \quad \frac{1}{2n} \sum_{k=1}^{r} \left\| y^{(k)} - X^{(k)} \left(s^{(k)} + b^{(k)} \right) \right\|_2^2 + \lambda_s \|S\|_{1,1} + \lambda_b \|B\|_{1,\infty} \quad (2.1)$$

Then output $\widehat{\Theta} = \widehat{B} + \widehat{S}$.

2.2.1 Our Method

Our method models the unknown parameter Θ as a superposition of a block-sparse matrix B (corresponding to the features shared across many tasks) and a sparse matrix S (corresponding to the features shared across few tasks). We estimate the sum of two parameter matrices B and S with different regularizations for each: encouraging block-structured row-sparsity in B and elementwise sparsity in S. The corresponding simple models would either just use block-sparse regularizations [101, 104] or just elementwise sparsity regularizations [131, 142], so that either method would perform better in certain suited regimes. Interestingly, as we will see in the main results, by explicitly allowing to have both block-sparse and elementwise sparse component (see Algorithm 2.2.1), we are able to *outperform both* classes of these "clean models", for *all* regimes $\bar{\Theta}$.

2.3 Main Results and Their Consequences

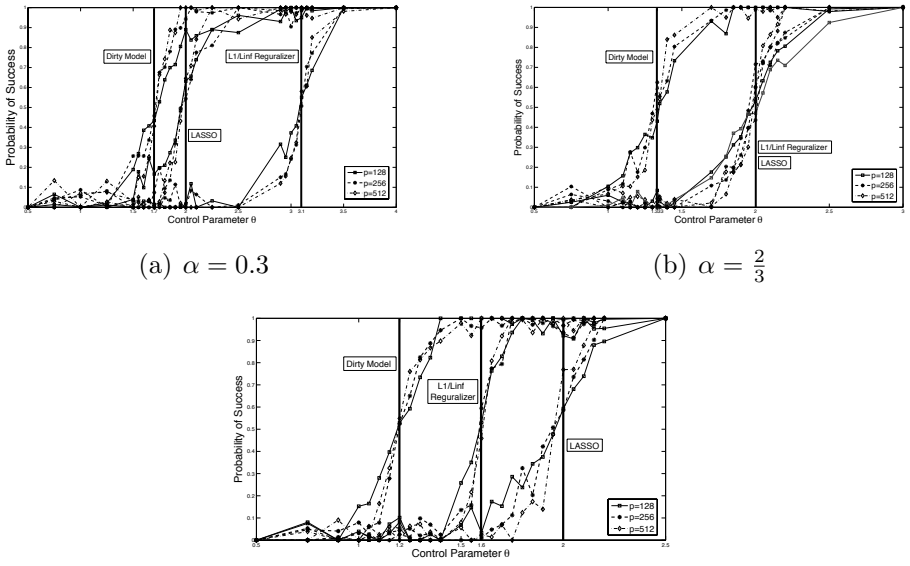
We now provide precise statements of our main results. A number of recent results have shown that the Lasso [131, 142] and ℓ_1/ℓ_∞ block-regularization [101] methods succeed in model selection, i.e., recovering signed supports with controlled error bounds under high-dimensional scaling regimes. Our first two theorems extend these results to our model setting. In Theorem 1, we consider the case of deterministic design matrices $X^{(k)}$, and provide sufficient conditions guaranteeing signed support recovery, and elementwise ℓ_∞ norm error bounds. In Theorem 2, we specialize this theorem to the case where the rows of the design matrices are random from a general zero mean Gaussian distribution: this allows us to provide scaling on the number of observations required in or-

der to guarantee signed support recovery and bounded elementwise ℓ_∞ norm error.

Our third result is the most interesting in that it explicitly quantifies the performance gains of our method vis-a-vis Lasso and the ℓ_1/ℓ_∞ block-regularization method. Since this entailed finding the precise constants underlying earlier theorems, and a correspondingly more delicate analysis, we follow [101] and focus on the case where there are two-tasks (i.e. $r = 2$), and where we have standard Gaussian design matrices as in Theorem 2. Further, while each of two tasks depends on s features, only a fraction α of these are common. It is then interesting to see how the behaviors of the different regularization methods vary with the extent of overlap α.

Comparisons. [101] show that there is actually a "phase transition" in the scaling of the probability of successful signed support-recovery with the number of observations. Denote a particular rescaling of the sample-size $\theta_{Lasso}(n, p, \alpha) = \frac{n}{s \log(p-s)}$. Then as [142] show, when the rescaled number of samples scales as $\theta_{Lasso} > 2 + \delta$ for any $\delta > 0$, Lasso succeeds in recovering the signed support of all columns with probability converging to one. But when the sample size scales as $\theta_{Lasso} < 2 - \delta$ for any $\delta > 0$, Lasso *fails* with probability converging to one. For the ℓ_1/ℓ_∞-regularized multiple linear regression, define a similar rescaled sample size $\theta_{1,\infty}(n, p, \alpha) = \frac{n}{s \log(p-(2-\alpha)s)}$. Then as [101] show there is again a transition in probability of success from near zero to near one, at the rescaled sample size of $\theta_{1,\infty} = (4 - 3\alpha)$. Thus, for $\alpha < 2/3$ ("less sharing") Lasso would perform better since its transition is at a smaller sample size, while for $\alpha > 2/3$ ("more sharing") the ℓ_1/ℓ_∞ regularized method would perform better.

As we show in our third theorem, the phase transition for our method occurs at the rescaled sample size of $\theta_{1,\infty} = (2 - \alpha)$, which is *strictly* before either the Lasso or the ℓ_1/ℓ_∞ regularized method except for the boundary cases: $\alpha = 0$, i.e. the case of no sharing, where we *match* Lasso, and for $\alpha = 1$, i.e. full sharing, where we *match* ℓ_1/ℓ_∞. Everywhere else, we *strictly outperform both* methods. Figure 2.3 shows the empirical performance of each of the three methods; as can be seen, they agree very well with the theoretical analysis. (Further details in the experiments Section 2.4).



(a) $\alpha = 0.3$ (b) $\alpha = \frac{2}{3}$

(c) $\alpha = 0.8$

Figure 2.1: Probability of success in recovering the true signed support using dirty model, Lasso and ℓ_1/ℓ_∞ regularizer. For a 2-task problem, the probability of success for different values of feature-overlap fraction α is plotted. As we can see in the regimes that Lasso is better than, as good as and worse than ℓ_1/ℓ_∞ regularizer ((a), (b) and (c) respectively), the dirty model outperforms both of the methods, i.e., it requires less number of observations for successful recovery of the true signed support compared to Lasso and ℓ_1/ℓ_∞ regularizer. Here $s = \lfloor \frac{p}{10} \rfloor$ always.

2.3.1 Sufficient Conditions for Deterministic Designs

We first consider the case where the design matrices $X^{(k)}$ for $k = 1, \cdots, r$ are deterministic, and start by specifying the assumptions we impose on the model. We note that similar sufficient conditions for the deterministic $X^{(k)}$'s case were imposed in papers analyzing Lasso [142] and block-regularization methods [101, 104].

A0 *Column Normalization:* $\|X_j^{(k)}\|_2 \leq \sqrt{2n}$ for all $j = 1, \ldots, p$ and $k = 1, \ldots, r$.

A1 *Incoherence Condition:*

$$\gamma_b := 1 - \max_{j \in \mathcal{U}^c} \sum_{k=1}^{r} \left\| \left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)} \left(\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \right\rangle \right\|_1 > 0,$$

where, $\mathcal{U}_k$ denotes the support of the k-th column of $\bar{\Theta}$, and $\mathcal{U} = \bigcup_k \mathcal{U}_k$ denotes the union of the supports of all tasks. We will also find it useful to define

$$\gamma_s := 1 - \max_{1 \le k \le r} \max_{j \in \mathcal{U}_k^c} \left\| \left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \left(\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \right\|_1 .$$

Note that by the incoherence condition **A1**, we have $\gamma_s > 0$.

A2 *Minimum Curvature Condition:*

$$C_{min} := \min_{1 \le k \le r} \lambda_{min} \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right) > 0.$$

Also, define $D_{max} := \max_{1 \le k \le r} \left\| \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \right\|_{\infty,1}$. As a consequence of **A2**, we have that $D_{\max}$ is finite.

A3 *Regularizers:* We require the regularization parameters satisfy

A3-1 $\lambda_s > \frac{2(2-\gamma_s)\sigma\sqrt{\log(pr)}}{\gamma_s \sqrt{n}}$.

A3-2 $\lambda_b > \frac{2(2-\gamma_b)\sigma\sqrt{\log(pr)}}{\gamma_b \sqrt{n}}$.

A3-3 $1 \le \frac{\lambda_b}{\lambda_s} \le r$ and $\frac{\lambda_b}{\lambda_s}$ is not an integer (see Lemma 11 and 12 for the reason).

Theorem 1. *Suppose* **A0-A3** *hold, and that we obtain estimate* $\widehat{\Theta}$ *from our algorithm. Then, with probability at least* $1 - c_1 \exp(-c_2 n)$*, we are guaranteed that the convex program (2.1) has a unique optimum and*

(a) The estimate $\widehat{\Theta}$ has no false inclusions, and has bounded ℓ_∞ norm error:

$$Supp(\widehat{\Theta}) \subseteq Supp(\bar{\Theta}), \quad \text{and}$$
$$\|\widehat{\Theta} - \bar{\Theta}\|_{\infty,\infty} \leq \underbrace{\sqrt{\frac{4\sigma^2 \log(pr)}{n\, C_{min}}} + \lambda_s D_{max}}_{b_{\min}} . \tag{2.2}$$

(b) The estimate $\widehat{\Theta}$ has no false exclusions, i.e., $sign(Supp(\widehat{\Theta})) = sign\left(Supp(\bar{\Theta})\right)$ provided that $\min_{(j,k)\in Supp(\bar{\Theta})} \left|\bar{\theta}_j^{(k)}\right| > b_{\min}$ for $b_{\min}$ defined in part (a).

The positive constants c_1, c_2 depend only on $\gamma_s, \gamma_b, \lambda_s, \lambda_b$ and σ, but are otherwise independent of n, p, r, the problem dimensions of interest.

Remark: Condition (a) guarantees that the estimate will have no *false inclusions*; i.e. all included features will be relevant. If in addition, we require that it have no *false exclusions* and that recover the support exactly, we need to impose the assumption in (b) that the non-zero elements are large enough to be detectable above the noise.

2.3.2 General Gaussian Designs

Often the design matrices consist of samples from a Gaussian ensemble (e.g. in Gaussian graphical model structure learning). Suppose that for each task $k = 1, \ldots, r$ the design matrix $X^{(k)} \in \mathbb{R}^{n\times p}$ is such that each row $X_i^{(k)} \in \mathbb{R}^p$ is a zero-mean Gaussian random vector with covariance matrix $\Sigma^{(k)} \in \mathbb{R}^{p\times p}$, and is independent of every other row. Let $\Sigma^{(k)}_{\mathcal{V},\mathcal{U}} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{U}|}$ be the submatrix of $\Sigma^{(k)}$ with corresponding rows to $\mathcal{V}$ and columns to $\mathcal{U}$. We require these covariance matrices to satisfy the following conditions:

C1 *Incoherence Condition:*

$$\gamma_b := 1 - \max_{j\in\mathcal{U}^c} \sum_{k=1}^{r} \left\| \Sigma^{(k)}_{j,\mathcal{U}_k}, \left(\Sigma^{(k)}_{\mathcal{U}_k,\mathcal{U}_k}\right)^{-1} \right\|_1 > 0 .$$

C2 *Minimum Curvature Condition:*

$$C_{min} := \min_{1 \le k \le r} \lambda_{min}\left(\Sigma^{(k)}_{\mathcal{U}_k,\mathcal{U}_k}\right) > 0$$

and let $D_{max} := \left\|\left(\Sigma^{(k)}_{\mathcal{U}_k,\mathcal{U}_k}\right)^{-1}\right\|_{\infty,1}$.

These conditions are analogues of the conditions for deterministic designs; they are now imposed on the covariance matrix of the (randomly generated) rows of the design matrix.

C3 *Regularizers:* Defining $s := \max_k |\mathcal{U}_k|$, we require the regularization parameters satisfy

C3-1 $\lambda_s \ge \frac{\left(4\sigma^2 C_{min}\log(pr)\right)^{1/2}}{\gamma_s\sqrt{nC_{min}} - \sqrt{2s\log(pr)}}$.

C3-2 $\lambda_b \ge \frac{\left(4\sigma^2 C_{min} r(r\log(2)+\log(p))\right)^{1/2}}{\gamma_b\sqrt{nC_{min}} - \sqrt{2sr(r\log(2)+\log(p))}}$.

C3-3 $1 \le \frac{\lambda_b}{\lambda_s} \le r$ and $\frac{\lambda_b}{\lambda_s}$ is not an integer.

Theorem 2. *Suppose assumptions* **C1-C3** *hold, and that the number of samples scale as*

$$n > \max\left(\frac{2s\log(pr)}{C_{min}\gamma_s^2}, \frac{2sr\big(r\log(2)+\log(p)\big)}{C_{min}\gamma_b^2}\right).$$

Suppose we obtain estimate $\widehat{\Theta}$ *from our algorithm. Then, with probability at least*

$$1 - c_1\exp\left(-c_2\left(r\log(2)+\log(p)\right)\right) - c_3\exp(-c_4\log(rs)) \to 1$$

for some positive numbers $c_1 - c_4$*, we are guaranteed that the algorithm estimate* $\widehat{\Theta}$ *is unique and satisfies the following conditions:*

(a) The estimate $\widehat{\Theta}$ has no false inclusions, and has bounded ℓ_∞ norm error so that

$$Supp(\widehat{\Theta}) \subseteq Supp(\bar{\Theta}), \quad \textit{and}$$
$$\|\widehat{\Theta} - \bar{\Theta}\|_{\infty,\infty} \leq \underbrace{\sqrt{\frac{50\sigma^2 \log(rs)}{nC_{min}}} + \lambda_s \left(\frac{4s}{C_{min}\sqrt{n}} + D_{max} \right)}_{g_{\min}}. \tag{2.3}$$

(b) The estimate $\widehat{\Theta}$ has no false exclusions, i.e., $sign(Supp(\widehat{\Theta})) = sign\left(Supp(\bar{\Theta})\right)$ provided that $\min_{(j,k)\in Supp(\bar{\Theta})} \left|\bar{\theta}_j^{(k)}\right| > g_{\min}$ for $g_{\min}$ defined in part (a).

2.3.3 Quantifying the gain for 2-Task Gaussian Designs

This is one of the most important results of this chapter. Here, we perform a more delicate and finer analysis to establish precise quantitative gains of our method. We focus on the special case where $r = 2$ and the design matrix has rows generated from the standard Gaussian distribution $\mathcal{N}(0, I_{n\times n})$. As we will see both analytically and experimentally, our method strictly outperforms both Lasso and ℓ_1/ℓ_∞-block-regularization over for all cases, except at the extreme endpoints of no support sharing (where it matches that of Lasso) and full support sharing (where it matches that of ℓ_1/ℓ_∞). We now present our analytical results; the empirical comparisons are presented next in Section 2.4. The results will be in terms of a particular rescaling of the sample size n as

$$\theta(n, p, s, \alpha) := \frac{n}{(2-\alpha)s \log\left(p - (2-\alpha)s\right)}.$$

We also require that the regularizers satisfy

$$\textbf{F1}\ \lambda_s > \frac{\left(4\sigma^2(1-\sqrt{s/n})(\log(r) + \log(p-(2-\alpha)s))\right)^{1/2}}{\sqrt{n} - \sqrt{s} - \left((2-\alpha)\ s\ (\log(r) + \log(p-(2-\alpha)s))\right)^{1/2}}.$$

$$\textbf{F2}\ \lambda_b > \frac{\left(4\sigma^2(1-\sqrt{s/n})r(r\log(2) + \log(p-(2-\alpha)s))\right)^{1/2}}{\sqrt{n} - \sqrt{s} - \left((1-\alpha/2)\ sr\ (r\log(2) + \log(p-(2-\alpha)s))\right)^{1/2}}.$$

F3 $\frac{\lambda_b}{\lambda_s} = \sqrt{2}$.

Theorem 3. *Consider a 2-task regression problem (n, p, s, α), where the design matrix has rows generated from the standard Gaussian distribution $\mathcal{N}(0, I_{n\times n})$. Suppose*

$$\max_{j\in B^*} \left| \left|\Theta_j^{*(1)}\right| - \left|\Theta_j^{*(2)}\right| \right| \leq c\lambda_s,$$

where, B^* is the submatrix of Θ^* with rows where both entries are non-zero and c is a constant specified in Lemma 7. Then the estimate $\widehat{\Theta}$ of the problem (2.1) satisfies the following:

(***Success***) *Suppose the regularization coefficients satisfy* **F1** $-$ **F3**. *Further, assume that the number of samples scales as $\theta(n, p, s, \alpha) > 1$. Then, with probability at least $1 - c_1 \exp(-c_2 n)$ for some positive numbers c_1 and c_2, we are guaranteed that $\widehat{\Theta}$ satisfies the support-recovery and ℓ_∞ error bound conditions (a-b) in Theorem 2.*

(***Failure***) *If $\theta(n, p, s, \alpha) < 1$ there is no solution $(\hat{B}, \hat{S})$ for any choices of λ_s and λ_b such that sign $\Big($Supp$(\widehat{\Theta})\Big)$ = sign $\big($Supp$(\bar{\Theta})\big)$.*

Remark: The assumption on the gap $\left| \left|\Theta_j^{*(1)}\right| - \left|\Theta_j^{*(2)}\right| \right| \leq c\lambda_s$ reflects the fact that we require that most values of Θ^* to be balanced on both tasks on the shared support. As we show in a more general theorem (Theorem 4) in Section 2.6.3, even in the case where the gap is large, the dependence of the sample scaling on the gap is quite weak.

2.4 Simulation Results

In this section, we provide some simulation results. First, using our synthetic data set, we investigate the consequences of Theorem 3 when we have $r = 2$ tasks to learn. As we see, the empirical result verifies our theoretical guarantees. Next, we apply our method regression to a real datasets: a handwritten digit classification dataset with $r = 10$ tasks (equal to the number of

digits $0-9$). For this dataset, we show that our method outperforms both LASSO and ℓ_1/ℓ_∞ practically. For each method, the parameters are chosen via cross-validation; see supplemental material for more details.

2.4.1 Synthetic Data Simulation

Consider a $r = 2$-task regression problem of the form (n, p, s, α) as discussed in Theorem 3. For a fixed set of parameters (n, s, p, α), we generate 100 instances of the problem. Then, we solve the same problem using our model, ℓ_1/ℓ_∞ regularizer and LASSO by searching for penalty regularizer coefficients independently for each one of these programs to find the best regularizer by cross validation. After solving the three problems, we compare the signed support of the solution with the true signed support and decide whether or not the program was successful in signed support recovery. We describe these process in more details in this section.

Data Generation: We explain how we generated the data for our simulation here. We pick three different values of $p = 128, 256, 512$ and let $s = \lfloor 0.1p \rfloor$. For different values of α, we let $n = c\, s \log(p - (2-\alpha)s)$ for different values of c. We generate a random sign matrix $\widetilde{\Theta}^* \in \mathbb{R}^{p\times 2}$ (each entry is either 0, 1 or -1) with column support size s and row support size $(2-\alpha)s$ as required by Theorem 3. Then, we multiply each row by a real random number with magnitude greater than the minimum required for sign support recovery by Theorem 3. We generate two sets of matrices $X^{(1)}$, $X^{(2)}$ and W and use one of them for training and the other one for cross validation (test), subscripted Tr and Ts, respectively. Each entry of the noise matrices $W_{\mathrm{Tr}}, W_{\mathrm{Ts}} \in \mathbb{R}^{n\times 2}$ is drawn independently according to $\mathcal{N}(0, \sigma^2)$ where $\sigma = 0.1$. Each row of a design matrix $X^{(k)}_{\mathrm{Tr}}, X^{(k)}_{\mathrm{Ts}} \in \mathbb{R}^{n\times p}$ is sampled, independent of any other rows, from $\mathcal{N}(0, \mathbf{I}_{2\times 2})$ for all $k = 1, 2$. Having $X^{(k)}$, $\bar{Theta}$ and W in hand, we can calculate $Y_{\mathrm{Tr}}, Y_{\mathrm{Ts}} \in \mathbb{R}^{n\times 2}$ using the model $y^{(k)} = X^{(k)}\theta^{(k)} + w^{(k)}$ for all $k = 1, 2$ for both train and test set of variables.

Coordinate Descent Algorithm: Given the generated data $X^{(k)}_{\mathrm{Tr}}$ for $k = 1, 2$ and Y_{Tr} in the previous section, we want to recover matrices $\hat{B}$

and $\hat{S}$ that satisfy (2.1). We use the coordinate descent algorithm to numerically solve the problem (see Appendix 2.8). The algorithm inputs the tuple $(X_{\text{Tr}}^{(1)}, X_{\text{Tr}}^{(2)}, Y_{\text{Tr}}, \lambda_s, \lambda_b, \epsilon, \underline{B}, \underline{S})$ and outputs a matrix pair $(\hat{B}, \hat{S})$. The inputs $(\underline{B}, \underline{S})$ are initial guess and can be set to zero. However, when we search for optimal penalty regularizer coefficients, we can use the result for previous set of coefficients (λ_b, λ_s) as a good initial guess for the next coefficients $(\lambda_b + \xi, \lambda_s + \zeta)$. The parameter ϵ captures the stopping criterion threshold of the algorithm. We iterate inside the algorithm until the relative update change of the objective function is less than ϵ. Since we do not run the algorithm completely (until $\epsilon = 0$ works), we need to filter the small magnitude values in the solution $(\hat{B}, \hat{S})$ and set them to be zero.

Choosing penalty regularizer coefficients: Dictated by optimality conditions, we have $1 > \frac{\lambda_s}{\lambda_b} > \frac{1}{2}$. Thus, searching range for one of the coefficients is bounded and known. We set $\lambda_b = c\sqrt{\frac{rlog(p)}{n}}$ and search for $c \in [0.01, 100]$, where this interval is partitioned logarithmic. For any pair (λ_b, λ_s) we compute the objective function of Y_{Ts} and $X_{\text{Ts}}^{(k)}$ for $k = 1, 2$ using the filtered $(\hat{B}, \hat{S})$ from the coordinate descent algorithm. Then across all choices of (λ_b, λ_s), we pick the one with minimum objective function on the test data. Finally we let $\hat{\Theta} = \text{Filter}(\hat{B} + \hat{S})$ for $(\hat{B}, \hat{S})$ corresponding to the optimal (λ_b, λ_s).

Performance Analysis: We ran the algorithm for five different values of the overlap ratio $\alpha \in \{0.3, \frac{2}{3}, 0.8\}$ with three different number of features $p \in \{128, 256, 512\}$. For any instance of the problem (n, p, s, α), if the recovered matrix $\hat{\Theta}$ has the same sign support as the true $\bar{\Theta}$, then we count it as success, otherwise failure (even if one element has different sign, we count it as failure).

As Theorem 3 predicts and Fig 2.3 shows, the right scaling for the number of oservations is $\frac{n}{s\log(p-(2-\alpha)s)}$, where all curves stack on the top of each other at $2 - \alpha$. Also, the number of observations required by our model for true signed support recovery is always less than both LASSO and ℓ_1/ℓ_∞ regularizer. Fig 2.1(a) shows the probability of success for the case $\alpha = 0.3$ (when LASSO is better than ℓ_1/ℓ_∞ regularizer) and that our model outperforms both

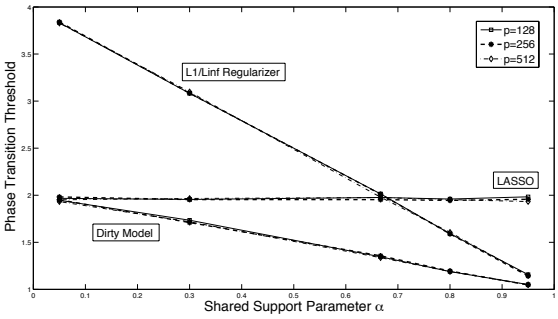


Figure 2.2: Verification of the result of the Theorem 3 on the behavior of phase transition threshold by changing the parameter α in a 2-task (n, p, s, α) problem for our method, LASSO and ℓ_1/ℓ_∞ regularizer. The y-axis is $\frac{n}{s\log(p-(2-\alpha)s)}$, where n is the number of samples at which threshold was observed. Here $s = \lfloor \frac{p}{10} \rfloor$. Our method shows a gain in sample complexity over the entire range of sharing α. The pre-constant in Theorem 3 is also validated.

methods. When $\alpha = \frac{2}{3}$ (see Fig 2.1(b)), LASSO and ℓ_1/ℓ_∞ regularizer performs the same; but our model require almost 33% less observations for the same performance. As α grows toward 1, e.g. $\alpha = 0.8$ as shown in Fig 2.1(c), ℓ_1/ℓ_∞ performs better than LASSO. Still, our model performs better than both methods in this case as well.

Scaling Verification: To verify that the phase transition threshold changes linearly with α as predicted by Theorem 3, we plot the phase transition threshold versus α. For five different values of $\alpha \in \{0.05, 0.3, \frac{2}{3}, 0.8, 0.95\}$ and three different values of $p \in \{128, 256, 512\}$, we find the phase transition threshold for our model, LASSO and ℓ_1/ℓ_∞ regularizer. We consider the point where the probability of success in recovery of signed support exceeds 50% as the phase transition threshold. We find this point by interpolation on the closest two points. Fig 2.2 shows that phase transition threshold for our model is always lower than the phase transition for LASSO and ℓ_1/ℓ_∞ regularizer.

	Feature	Size	Type	Dynamic Range
1	Pixel Shape (15×16)	240	Integer	0-6
2	2D Fourier Transform Coefficients	74	Real	0-1
3	Karhunen-Loeve Transform Coeficients	64	Real	-17:17
4	Profile Correlation	216	Integer	0-1400
5	Zernike Moments	46	Real	0-800
6	Morphological Features	3	Integer	0-6
		1	Real	100-200
		1	Real	1-3
		1	Real	1500-18000

Table 2.1: Six different classes of features provided in the dataset. The dynamic ranges are approximate not exact. The dynamic range of different morphological features are completely different. For those 6 morphological features, we provide their different dynamic ranges separately.

2.4.2 Handwritten Digits Dataset

We use a handwritten digit dataset to illustrate the performance of our method. According to the description of the dataset, this dataset consists of features of handwritten numerals (0-9) extracted from a collection of Dutch utility maps [44]. This dataset has been used by a number of papers [60, 137] as a reliable dataset for handwritten recognition algorithms.

Structure of the Dataset: In this dataset, there are 200 instances of handwritten digits 0-9 (totally 2000 digits). Each instance of each digit is scanned to an image of the size 30×48 pixels. This image is NOT provided by the dataset. Using the full resolution image of each digit, the dataset provides six different classes of features. A total of 649 features are provided for each instance of each digit. The information about each class of features is provided in Table 2.1. The combined handwriting images of the record number 100 is shown in Fig 2.3 (ten images are concatenated together with a spacer between each two).

Fitting the dataset to our model: Regardless of the nature of the features, we have 649 features for each of 200 instance of each digit. We need

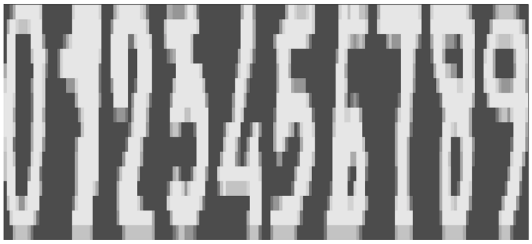


Figure 2.3: An instance of images of the ten digits extracted from the dataset

to learn $K = 10$ different tasks corresponding to ten different digits. To make the associated numbers of features comparable, we shrink the dynamic range of each feature to the interval -1 and 1. We divide each feature by an appropriate number (perhaps larger than the maximum of that feature in the dataset) to make sure that the dynamic range of all features is a (not too small) subset of $[-1, 1]$. Notice that in this division process, we don't care about the minimum and maximum of the training set. We just divide each feature by a fixed and predetermined number we provided as maximum in Table 2.1. For example, we divide the Pixel Shape feature by 6, Karhunen-Loeve coefficients by 17 or the last morphological feature by 18000 and so on. We do not shift the data; we only scale it.

Out of 200 samples provided for each digit, we take $n \leq 200$ samples for training. Let $X^{(k)} = X \in \mathbb{R}^{10n \times 649}$ for all $0 \leq k \leq 9$ be the matrix whose first n rows correspond to the features of the digit 0, the second n rows correspond to the features of the digit 1 and so on. Consequently, we set the vector $y^{(k)} \in \{0, 1\}^{10n}$ to be the vector such that $y_j^{(k)} = 1$ if and only if the j^{th} row of the feature matrix X corresponds to the digit k. This setup is called binary classification setup.

We want to find a block-sparse matrix $\hat{B} \in \mathbb{R}^{649 \times 10}$ and a sparse matrix $\hat{S} \in \mathbb{R}^{649 \times 10}$, so that for a given feature vector $\mathbf{x} \in \mathbb{R}^{649}$ extracted from the image of a handwritten digit $0 \leq k^* \leq 9$, we ideally have $k^* = \arg\max_{0 \leq k \leq 9} \mathbf{x} \left(\hat{B} + \hat{S} \right)$.

$\frac{n}{200}$		Our Model		ℓ_1/ℓ_∞	LASSO
5%	Average Classification Error	8.6%		9.9%	10.8%
	Variance of Error	0.53%		0.64%	0.51%
	Average Row Support Size	B:165	$B+S$:171	170	123
	Average Support Size	S:18	$B+S$:1651	1700	539
10%	Average Classification Error	3.0%		3.5%	4.1%
	Variance of Error	0.56%		0.62%	0.68%
	Average Row Support Size	B:211	$B+S$:226	217	173
	Average Support Size	S:34	$B+S$:2118	2165	821
20%	Average Classification Error	2.2%		3.2%	2.8%
	Variance of Error	0.57%		0.68%	0.85%
	Average Row Support Size	B:270	$B+S$:299	368	354
	Average Support Size	S:67	$B+S$:2761	3669	2053

Table 2.2: Simulation Results for our model, ℓ_1/ℓ_∞ and LASSO.

To find such matrices $\hat{B}$ and $\hat{S}$, we solve (2.1). We tune the parameters λ_b and λ_s in order to get the best result by cross validation. Since we have 10 tasks, we search for $\frac{\lambda_s}{\lambda_b} \in \left[\frac{1}{10}, 1\right]$ and let $\lambda_b = c\sqrt{\frac{2log(649)}{n}} \approx \frac{5c}{\sqrt{n}}$, where, empirically $c \in [0.01, 10]$ is a constant to be searched.

Performance Analysis: Table 2.2 shows the results of our analysis for different sizes of the training set as $\frac{n}{200}$. We measure the classification error on the test set for each digit to get the 10-vector of errors. Then, we find the average error and the variance of the error vector to show how the error is distributed over all tasks. We compare our method with ℓ_1/ℓ_∞ reguralizer method and LASSO.

2.5 Proof Outline

In this section we illustrate the proof outline of all three theorems as they are very similar in the nature. First, we introduce some notations and definitions and then, we provide a three step proof technique that we used to prove all three theorems.

2.5.1 Definitions and Setup

In this section, we rigorously define the terms and notation we used throughout the proofs.

Notation: For a vector v, the norms ℓ_1, ℓ_2 and ℓ_∞ are denoted as $\|v\|_1 = \sum_k \left|v^{(k)}\right|$, $\|v\|_2 = \sqrt{\sum_k \left|v^{(k)}\right|^2}$ and $\|v\|_\infty = \max_k \left|v^{(k)}\right|$, respectively. Also, for a matrix $Q \in \mathbb{R}^{p\times r}$, the norm ℓ_ζ/ℓ_ρ is denoted as $\|Q\|_{\rho,\zeta} = \|\left(\|q_1\|_\zeta, \cdots, \|q_p\|_\zeta\right)\|_\rho$. The maximum singular value of Q is denoted as $\lambda_{max}(Q)$. For a matrix $X \in \mathbb{R}^{n\times p}$ and a set of indices $\mathcal{U} \subseteq \{1, \cdots, p\}$, the matrix $X_\mathcal{U} \in \mathbb{R}^{n\times|\mathcal{U}|}$ represents the sub-matrix of X consisting of X_j's where $j \in \mathcal{U}$.

2.5.1.1 Towards Identifying Optimal Solution

This is a key step in our analysis. Our proof proceeds by choosing a pair $\widehat{B}, \widehat{S}$ such that the signed support of $\widehat{B} + \widehat{S}$ is the same as that of $\bar{\Theta}$, and then certifying that, under our assumptions, this pair is the optimum of the optimization problem (2.1). We construct this pair via a surrogate optimization problem – dubbed *oracle problem* in the literature as well as our proof outline below – which adds extra constraints to (2.1) in a way that ensures signed support recovery. Making the oracle problem is a key step in our proof.

For (2.1), let $d = \lceil \frac{\lambda_b}{\lambda_s} \rceil$; in this paper we will always have $1 \leq d \leq r$, where we recall r is the number of tasks. Using this d, we now define two matrices B^*, S^*, such that $B^* + S^* = \bar{\Theta}$, as follows. In each row $\bar{\Theta}_j$, let v_j be the $(d+1)^{th}$ largest magnitude of the elements in Θ_j. Then, the $(j,k)^{th}$ element $s_j^{*(k)}$ of the matrix S^* is defined as follows

$$s_j^{*(k)} \;=\; \text{sign}(\theta_j^{(k)}) \max\left\{0, \left|\theta_j^{(k)}\right| - v_j\right\}$$

In words, to obtain S^* we take the matrix $\bar{\Theta}$ and for each element we *clip its magnitude* to be the *excess* over the $(d+1)^{th}$ largest magnitude in its row. We retain the sign. Finally, define $B^* = \bar{\Theta} - S^*$ to be the residual. It is thus clear that

- S^* will have at most d non-zero elements in each row.
- Each row of B^* is either identically 0, or has at least d non-zero elements. Also, in the latter case, at least d of them have the same magnitude.
- If any element (j,k) is non-zero in both S^* and B^* then its sign is the same in both.

S^* thus takes on the role of the "true sparse matrix", and B^* the role of the "true block-sparse matrix". We will use B^*, S^* to construct our oracle problem later. The pair also has the following significance: our results will imply that if we have infinite samples, then B^*, S^* will be the solution to (2.1).

2.5.1.2 Sparse Matrix Setup

For any matrix S, define $\text{Supp}(S) = \{(j,k) : s_j^{(k)} \neq 0\}$, and let $U_s = \{S \in \mathbb{R}^{p\times r} : \text{Supp}(S) \subseteq \text{Supp}(S^*)\}$ be the subspace of matrices whose their support is the subset of the matrix S^*. The orthogonal projection to the subspace U_s can be defined as follows:

$$\left(P_{U_s}(S)\right)_{j,k} = \begin{cases} s_j^{(k)} & (j,k) \in \text{Supp}(S^*) \\ 0 & \text{ow.} \end{cases}$$

We can define the orthogonal complement space of U_s to be $U_s^c = \{S \in \mathbb{R}^{p\times r} : \text{Supp}(S) \cap \text{Supp}(S^*) = \phi\}$. The orthogonal projection to this space can be defined as $P_{U_s^c}(S) = S - P_{U_s}(S)$. Since the type of the block-sparsity we consider is a block-sparsity assumption on the rows of matrices, we need to characterize the sparsity of the rows of the matrix S^*. This motivates to define $D(S) = \max_{1\leq j\leq p} \|s_j\|_0$ denoting the maximum number of non-zero elements in any row of the sparse matrix S.

2.5.1.3 Row-Sparse Matrix Setup

For any matrix B, define $\text{RowSupp}(B) = \{j : \exists k \text{ s.t. } b_j^{(k)} \neq 0\}$, and let $U_b = \{B \in \mathbb{R}^{p\times r} : \text{RowSupp}(B) \subseteq \text{RowSupp}(B^*)\}$ be the subspace of matrices

whose their row support is the subset of the row support of the matrix B^*. The orthogonal projection to the subspace U_b can be defined as follows:

$$\left(P_{U_b}(B)\right)_j = \begin{cases} b_j & j \in \text{RowSupp}(B^*) \\ \mathbf{0} & \text{ow.} \end{cases}$$

We can define the orthogonal complement space of U_b to be $U_b^c = \{B \in \mathbb{R}^{p\times r} : \text{RowSupp}(B) \cap \text{RowSupp}(B^*) = \phi\}$. The orthogonal projection to this space can be defined as $P_{U_b^c}(B) = B - P_{U_b}(B)$.

For a given matrix $B \in \mathbb{R}^{p\times r}$, let $M_j(B) = \{k : |b_j^{(k)}| = \|b_j\|_\infty > 0\}$ be the set of indices that the corresponding elements achieve the maximum magnitude on the j^{th} row with positive or negative signs. Also, let $M(B) = \min_{1\le j\le p} |M_j(B)|$ be the minimum number of elements who achieve the maximum in each row of the matrix B.

The following technical lemma is useful in the proof of all three theorems.

Lemma 1. *If* $(B, S) = \mathcal{H}_d(\Theta)$ *then*

(P1) $M(B) \ge d + 1$ *and* $D(S) \le d$.

(P2) $sign(s_j^{(k)}) = sign(b_j^{(k)})$ *for all* $j \in RowSupp(B)$ *and* $k \in M_j(B)$.

(P3) $s_j^{(k)} = 0$ *for all* $j \in RowSupp(B)$ *and* $k \notin M_j(B)$.

Proof. The proof follows from the definition of $\mathcal{H}$.

□

2.5.2 Proof Overview

The proofs of all three of our theorems follow a primal-dual witness technique, and consist of two steps, as detailed in this section. The first step constructs a primal-dual witness candidate, and is common to all three theorems. The second step consists of showing that the candidate constructed in

the first step is indeed a primal-dual witness. The theorem proofs differ in this second step, and show that under the respective conditions imposed in the theorems, the construction succeeds with high probability. These steps are as follows:

STEP 1: Denote the true optimal solution pair $(B^*, S^*) = \mathcal{H}_d(\bar{\Theta})$ as defined in Section 2.5.1.1, for $d = \lfloor \frac{\lambda_b}{\lambda_s} \rfloor$. See Lemma 1 for basic properties of these matrices B^* and S^*.

Primal Candidate: We can then design a candidate optimal solution $(\tilde{S}, \tilde{B})$ with the desired sparsity pattern using a restricted support optimization problem, called *oracle problem*:

$$(\tilde{S}, \tilde{B}) \in \arg \min_{S \in U_s, B \in U_b} \frac{1}{2n} \sum_{k=1}^{r} \left\| y^{(k)} - X^{(k)} \left(s^{(k)} + b^{(k)} \right) \right\|_2^2 + \lambda_s \|S\|_{1,1} + \lambda_b \|B\|_{1,\infty}. \tag{2.4}$$

Dual Candidate: We set $\widetilde{Z}_{\bigcup_{k=1}^r \mathcal{U}_k}$ as the subgradient of the optimal primal parameters of (2.4) . Specifically, we set

$$\widetilde{Z}_{\bigcup_{k=1}^r \mathcal{U}_k} = \left(\widetilde{Z}_s \right)_{\bigcup_{k=1}^r \mathcal{U}_k} + \left(\widetilde{Z}_b \right)_{\bigcup_{k=1}^r \mathcal{U}_k},$$

where, $\widetilde{Z}_s = \lambda_s \text{sign}(\tilde{S})$, and for all $j \in \bigcup_{k=1}^r \mathcal{U}_k$,

$$(\tilde{z}_b)_j^{(k)} = \begin{cases} \dfrac{\lambda_b - \lambda_s \|\tilde{s}_j\|_0}{\left| M_j(\tilde{B}) \right| - \|\tilde{s}_j\|_0} \text{sign}\left(\tilde{b}_j^{(k)} \right) \\ \qquad\qquad k \in M_j(\tilde{B}) \quad \& \quad (j,k) \notin \text{Supp}(\tilde{S}) \cdot \\ 0 \qquad\qquad \text{ow} \end{cases}$$

To get an explicit form for $\widetilde{Z}_{\bigcap_{k=1}^r \mathcal{U}_k^c}$, let $\Delta = \tilde{B} + \tilde{S} - B^* - S^*$. From the optimality conditions for the oracle problem (2.4), we have

$$\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \Delta_{\mathcal{U}_k}^{(k)} - \frac{1}{n} \left(X_{\mathcal{U}_k}^{(k)} \right)^T w^{(k)} + \tilde{z}_{\mathcal{U}_k}^{(k)} = 0.$$

and consequently,

$$\Delta_{\mathcal{U}_k}^{(k)} = \left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\left(\frac{1}{n}\left(X_{\mathcal{U}_k}^{(k)}\right)^T w^{(k)} - \tilde{z}_{\mathcal{U}_k}^{(k)}\right). \tag{2.5}$$

Solving for $\tilde{z}^{(k)}_{\bigcap_{k=1}^r \mathcal{U}_k^c}$, for all $j \in \bigcap_{k=1}^r \mathcal{U}_k^c$, we get

$$\tilde{z}_j^{(k)} = -\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle \Delta_{\mathcal{U}_k}^{(k)} + \frac{1}{n}\left(X_j^{(k)}\right)^T w^{(k)}.$$

Substituting for the value of $\Delta_{\mathcal{U}_k}^{(k)}$, we get

$$\begin{aligned}\tilde{z}_j^{(k)} = \frac{1}{n}\left(X_j^{(k)}\right)^T w^{(k)} - \frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle \left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1} \\ \left(\frac{1}{n}\left(X_{\mathcal{U}_k}^{(k)}\right)^T w^{(k)} - \tilde{z}_{\mathcal{U}_k}^{(k)}\right).\end{aligned} \tag{2.6}$$

STEP 2: This step consists of showing that the pair $(\tilde{S}, \tilde{B}, \widetilde{Z})$ constructed in the earlier step is actually a *feasible* primal-dual pair of (2.1). This would then the required support-recovery result since the constructed primal candidate $\tilde{S}, \tilde{B}$ had the required sparsity pattern by *construction.*

We will make use of the following lemma that specifies a set of sufficient (stationary) optimality conditions for the $(\tilde{S}, \tilde{B})$ from (2.4) to be the unique solution of the (unrestricted) optimization problem (2.1):

Lemma 2. *Under our (stationary) assumptions on the design matrices $X^{(k)}$, the matrix pair $(\tilde{S}, \tilde{B})$ is the unique solution of the problem* (2.1) *if there exists a matrix $\widetilde{Z} \in \mathbb{R}^{p\times r}$ such that*

(C1) $P_{U_s}(\widetilde{Z}) = \lambda_s sign\left(\tilde{S}\right)$.

(C2) $P_{U_b}(\widetilde{Z}) = \begin{cases} t_j^{(k)}\, sign\left(\tilde{b}_j^{(k)}\right), & k \in M_j(B^*) \\ 0 & o.w.. \end{cases}$, *where,* $t_j^{(k)} \geq 0$ *such that* $\sum_{k\in M_j(B^*)} t_j^{(k)} = \lambda_b$.

(C3) $\left\|P_{U_s^c}(\widetilde{Z})\right\|_{\infty,\infty} < \lambda_s$.

(C4) $\left\|P_{U_b^c}(\widetilde{Z})\right\|_{\infty,1} < \lambda_b$.

(C5) $\frac{1}{n}\left\langle X^{(k)}, X^{(k)}\right\rangle \left(\tilde{b}^{(k)}+\tilde{s}^{(k)}\right)-\frac{1}{n}(X^{(k)})^T y^{(k)}+\tilde{z}^{(k)}=0 \quad \forall 1 \leq k \leq r.$

Proof. By assumptions (C1) and (C3), $\frac{1}{\lambda_s}\widetilde{Z} \in \partial\|\tilde{S}\|_{1,1}$ and by assumptions (C2) and (C4), $\frac{1}{\lambda_b}\widetilde{Z} \in \partial\|\tilde{B}\|_{1,\infty}$. Thus, $(\tilde{S}, \tilde{B}, \widetilde{Z})$ is a feasible primal-dual pair of (2.1) according to the Lemma 36.

Let $\mathbb{B}$ and $\mathbb{S}$ to be balls of ℓ_∞/ℓ_1 and ℓ_∞/ℓ_∞ with radiuses λ_b and λ_s, respectively. Considering the fact that $\lambda_b\|B\|_{1,\infty} = \sup_{Z\in\mathbb{B}}\langle Z, B\rangle$ and $\lambda_s\|S\|_{1,1} = \sup_{Z\in\mathbb{S}}\langle Z, S\rangle$, the problem (2.1) can be written as

$$(\hat{S}, \hat{B}) = \arg\inf_{S,B}\sup_{Z\in\mathbb{B}\cap\mathbb{S}}\left\{\frac{1}{2n}\sum_{k=1}^{r}\left\|y^{(k)}-X^{(k)}\left(b^{(k)}+s^{(k)}\right)\right\|_2^2 + \langle Z, S\rangle + \langle Z, B\rangle\right\}.$$

This saddle-point problem is strictly feasible and convex-concave. Given any dual variable, in particular $\widetilde{Z}$, and any primal optimal $(\hat{S}, \hat{B})$ we have $\lambda_b\|\hat{B}\|_{1,\infty} = \left\langle \widetilde{Z}, \hat{B}\right\rangle$ and $\lambda_s\|\hat{S}\|_{1,1} = \left\langle \widetilde{Z}, \hat{S}\right\rangle$. This implies that $\hat{b}_j = \mathbf{0}$ if $\|\tilde{z}_j\|_1 < \lambda_b$ (because $\lambda_b\sum_j\|\hat{b}_j\|_\infty \leq \sum_j\|\tilde{z}_j\|_1\|\hat{b}_j\|_\infty$ and if $\|\tilde{z}_{j_0}\|_1 < \lambda_b$ for some j_0, then others can not compensate for that in the sum due to the fact that $\widetilde{Z} \in \mathbb{B}$, i.e., $\|\tilde{z}_j\|_1 \leq \lambda_b$). It also implies that $\hat{s}_j^{(k)} = 0$ if $\left|\tilde{z}_j^{(k)}\right| < \lambda_s$ for a similar reason. Hence, $P_{U_b^c}(\hat{B}) = 0$ and $P_{U_s^c}(\hat{S}) = 0$. This means that solving the restricted problem (2.4) is equivalent to solving the problem (2.1).

The uniqueness follows from our (stationary) assumptions on design matrices $X^{(k)}$ that the matrix $\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle$ is invertible for all $1 \leq k \leq r$.

Using this assumption, the problem (2.4) is *strictly* convex and the solution is unique. Consequently, the solution of (2.1) is also unique, since we showed that these two problems are equivalent. This concludes the proof of the lemma. □

By construction, the primal-dual pair $(\tilde{B}, \tilde{S}, \widetilde{Z})$ satisfies the (C1), (C2) and (C5) conditions in Lemma 33. *It only remains to guarantee (C3) and (C4) separately for each of the theorems.*

Indeed, this is where the proofs of the theorems differ. Specifically, Lemmas 3, 5 and 8 ensure these conditions are satisfied with given sample complexities in Theorems 1, 2 and 3, respectively.

2.6 Proofs

The proofs of our three main theorems are in sections 2.6.1, 2.6.2 and 2.6.3 respectively.

2.6.1 Proof of Theorem 1

Let $d = \lfloor \frac{\lambda_b}{\lambda_s} \rfloor$ and $(B^*, S^*) = \mathcal{H}_d(\bar{\Theta})$. Then, the result follows from Proposition 1 below.

Proposition 1 (Structure Recovery)**.** *Under assumptions of Theorem 1, with probability* $1 - c_1 \exp(-c_2 n)$ *for some positive constants* c_1 *and* c_2*, we are guaranteed that the following properties hold:*

(P1) Problem (2.1) *has unique solution* $(\hat{S}, \hat{B})$ *such that* $Supp(\hat{S}) \subseteq Supp(S^*)$ *and* $RowSupp(\hat{B}) \subseteq RowSupp(B^*)$*.*

(P2) $\left\| \hat{B} + \hat{S} - B^* - S^* \right\|_\infty \leq \underbrace{\sqrt{\frac{4\sigma^2 \log(pr)}{C_{min} n}} + \lambda_s D_{max}}_{b_{\min}}.$

(P3) $sign\left(Supp(\hat{s}_j)\right) = sign\left(Supp(s_j^*)\right)$
for all $j \notin RowSupp(B^*)$ *provided that*

$$\min_{\substack{j \notin RowSupp(B^*) \\ (j,k) \in Supp(S^*)}} \left| s_j^{*(k)} \right| > b_{\min}.$$

(P4) $sign\Big(Supp(\hat{s}_j + \hat{b}_j)\Big) = sign\left(Supp(s_j^* + b_j^*)\right)$
for all $j \in RowSupp(B^*)$ *provided that*

$$\min_{(j,k) \in Supp(B^*)} \left| b_j^{*(k)} + s_j^{*(k)} \right| > b_{\min}.$$

Proof. We prove the result separately for each part.

(P1) Considering the constructed primal-dual pair, it suffices to show that (C3) and (C4) in Lemma 33 are satisfied with high probability. By Lemma 3, with probability at least $1 - c_1 \exp(-c_2 n)$ those two conditions hold and hence, $(\hat{S}, \hat{B}) = (\tilde{S}, \tilde{B})$ is the unique solution of (2.1) and the property (P1) follows.

(P2) Using (2.5), we have

$$\begin{aligned}
\max_{j \in \mathcal{U}_k} \left| \Delta_j^{(k)} \right| &\leq \left\| \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \frac{1}{n} \left(X_{\mathcal{U}_k}^{(k)} \right)^T w^{(k)} \right\|_\infty \\
&\qquad + \left\| \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \tilde{z}_{\mathcal{U}_k}^{(k)} \right\|_\infty \\
&\leq \sqrt{\frac{4\sigma^2 \log(pr)}{C_{min} n}} + \lambda_s D_{max},
\end{aligned}$$

where, the second inequality holds with high probability as a result of Lemma 4 for $\alpha = \epsilon \sqrt{\frac{4\sigma^2 \log(pr)}{C_{min} n}}$ for some $\epsilon > 1$, considering the fact that

$\text{Var}\left(\Delta_j^{(k)}\right) \leq \frac{\sigma^2}{C_{min}n}$.

(P3) Using (P1) in Lemma 11, this event is equivalent to the event that for all $j \notin \text{RowSupp}(B^*)$ with $(j,k) \in \text{Supp}(S^*)$, we have $\left(\Delta_j^{(k)} + s_j^{*(k)}\right) \text{sign}\left(s_j^{*(k)}\right) > 0$. By Hoeffding inequality, we have

$$\begin{aligned}
&\mathbb{P}\left[\left(\Delta_j^{(k)} + s_j^{*(k)}\right) \text{sign}\left(s_j^{*(k)}\right) > 0\right] \\
&\qquad = \mathbb{P}\left[-\Delta_j^{(k)} \text{sign}\left(s_j^{*(k)}\right) < \left|s_j^{*(k)}\right| \right] \\
&\qquad \geq \mathbb{P}\left[\left|\Delta_j^{(k)}\right| < \left|s_j^{*(k)}\right| \right].
\end{aligned}$$

By part (P2), this event happens with high probability if $\min_{\substack{j \notin \text{RowSupp}(B^*) \\ (j,k) \in \text{Supp}(S^*)}} \left|s_j^{*(k)}\right| > b_{\min}$.

(P4) Using (P1) in Lemma 11, this event is equivalent to the event that for all $j \in \text{RowSupp}(B^*)$, we have $\left(\Delta_j^{(k)} + b_j^{*(k)} + s_j^{*(k)}\right) \text{sign}\left(b_j^{*(k)} + s_j^{*(k)}\right) > 0$. By Hoeffding inequality, we have

$$\begin{aligned}
&\mathbb{P}\left[\left(\Delta_j^{(k)} + b_j^{*(k)} + s_j^{*(k)}\right) \text{sign}\left(b_j^{*(k)} + s_j^{*(k)}\right) > 0\right] \\
&\qquad = \mathbb{P}\left[-\Delta_j^{(k)} \text{sign}\left(b_j^{*(k)} + s_j^{*(k)}\right) < \left|b_j^{*(k)} + s_j^{*(k)}\right| \right] \\
&\qquad \geq \mathbb{P}\left[\left|\Delta_j^{(k)}\right| < \left|b_j^{*(k)} + s_j^{*(k)}\right| \right].
\end{aligned}$$

By part (P2), this event happens with high probability if $\min_{(j,k) \in \text{Supp}(B^*)} \left|b_j^{*(k)} + s_j^{*(k)}\right| > b_{\min}$.

□

Lemma 3. *Under conditions of Proposition 1, the conditions (C3) and (C4) in Lemma 33 hold for the constructed primal-dual pair with probability at least* $1 - c_1 \exp(-c_2 n)$ *for some positive constants* c_1 *and* c_2.

Proof. First, we need to bound the projection of $\widetilde{Z}$ into the space U_s^c. Notice that

$$\left|\left(P_{U_s^c}(\widetilde{Z})\right)_j^{(k)}\right| = \begin{cases} \dfrac{\lambda_b - \lambda_s \|\tilde{s}_j\|_0}{\left|M_j(\tilde{B})\right| - \|\tilde{s}_j\|_0} \\ \quad j \in \text{RowSupp}(\tilde{B}) \;\&\; (j,k) \notin \text{Supp}(\tilde{S}) \\ \left|\tilde{z}_j^{(k)}\right| \qquad\qquad j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c \\ 0 \qquad\qquad \text{ow.} \end{cases}.$$

By our assumption on the ratio of the penalty regularizer coefficients, we have $\frac{\lambda_b - \lambda_s \|\tilde{s}_j\|_0}{\left|M_j(\tilde{B})\right| - \|\tilde{s}_j\|_0} < \lambda_s$. Moreover, we have

$$\begin{aligned} \left|\tilde{z}_j^{(k)}\right| &\le \max_{j \in \bigcap_{k=1}^r \mathcal{U}_k^c} \left\| \frac{1}{n} \left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \right\|_1 \\ &\qquad \left(\left\| \frac{1}{n} \left(X^{(k)}\right)^T w^{(k)} \right\|_\infty + \left\| \tilde{z}_{\mathcal{U}_k}^{(k)} \right\|_\infty \right) \\ &\qquad + \left\| \frac{1}{n} \left(X^{(k)}\right)^T w^{(k)} \right\|_\infty \\ &\le (2 - \gamma_s) \left\| \frac{1}{n} \left(X^{(k)}\right)^T w^{(k)} \right\|_\infty + (1 - \gamma_s) \left\| \tilde{z}_{\mathcal{U}_k}^{(k)} \right\|_\infty \\ &\le (2 - \gamma_s) \left\| \frac{1}{n} \left(X^{(k)}\right)^T w^{(k)} \right\|_\infty + (1 - \gamma_s) \lambda_s. \end{aligned}$$

Thus, the event $\|P_{U_s^c}(\widetilde{Z})\|_{\infty,\infty} < \lambda_s$ is equivalent to the event $\max_{1 \le k \le r} \left\| \frac{1}{n} \left(X^{(k)}\right)^T w^{(k)} \right\|_\infty < \frac{\gamma_s}{2 - \gamma_s} \lambda_s$. By Lemma 4, this event happens with probability at least $1 - 2 \exp\left(- \frac{\gamma_s^2 n \lambda_s^2}{4(2-\gamma_s)^2 \sigma^2} + \log(pr) \right)$. This probability goes to 1 if $\lambda_s > \frac{2(2-\gamma_s)\sigma\sqrt{\log(pr)}}{\gamma_s \sqrt{n}}$

as stated in the assumptions.

Next, we need to bound the projection of $\widetilde{Z}$ into the space U_b^c. Notice that

$$\sum_{k=1}^{r}\left|\left(P_{U_b^c}(\widetilde{Z})\right)_j^{(k)}\right| = \begin{cases} \lambda_s\|\tilde{s}_j\|_0 & j \in \bigcup_{k=1}^{r}\mathcal{U}_k - \text{RowSupp}(B^*) \\ \sum_{k=1}^{r}\left|\tilde{z}_j^{(k)}\right| & j \in \bigcap_{k=1}^{r}\mathcal{U}_k^c \\ 0 & \text{ow} \end{cases}.$$

We have $\lambda_s\|\tilde{s}_j\|_0 \leq \lambda_s D(S^*) < \lambda_b$ by our assumption on the ratio of the penalty regularizer coefficients. We can establish the following bound:

$$\begin{aligned}
&\sum_{k=1}^{r}\left|\tilde{z}_j^{(k)}\right| \\
&\leq \max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left\|\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\|_1 \\
&\qquad\qquad \left(\max_{j\in\bigcup_{k=1}^{r}\mathcal{U}_k}\left\|\tilde{z}_j^{(k)}\right\|_1 + \max_{1\leq k\leq r}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty\right) \\
&\qquad + \max_{1\leq k\leq r}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty \\
&\leq (1-\gamma_b)\lambda_b + (2-\gamma_b)\max_{1\leq k\leq K}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty.
\end{aligned}$$

Thus, the event $\|P_{U_b^c}(\widetilde{Z})\|_{\infty,1} < \lambda_b$ is equivalent to the event $\max_{1\leq k\leq r}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty < \frac{\gamma_b}{2-\gamma_b}\lambda_b$. By Lemma 4, this event happens with probability at least $1-2\exp\left(-\frac{\gamma_b^2 n\lambda_b^2}{4(2-\gamma_b)^2\sigma^2} + \log(pr)\right)$. This probability goes to 1 if $\lambda_b > \frac{2(2-\gamma_b)\sigma\sqrt{\log(pr)}}{\gamma_b\sqrt{n}}$ as stated in the assumptions.

Hence, with probability at least $1 - c_1\exp(-c_2 n)$ conditions (C3) and (C4) in Lemma 33 are satisfied.

□

Lemma 4.

$$\mathbb{P}\left[\max_{1\leq k\leq r}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty < \alpha\right] \geq 1 - 2\exp\left(-\frac{\alpha^2 n}{4\sigma^2} + \log(pr)\right).$$

Proof. Since $w_j^{(k)}$'s are distributed as $\mathcal{N}(0,\sigma^2)$, we have $\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}$ distributed as $\mathcal{N}\left(0, \frac{\sigma^2}{n}\left(X^{(k)}\right)^T X^{(k)}_{\mathcal{U}_k}\right)$. Using Hoeffding inequality, we have

$$\begin{aligned}
\mathbb{P}\left[\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty \geq \alpha\right] &\leq \sum_{j=1}^{p}\mathbb{P}\left[\left|\frac{1}{n}\left(X_j^{(k)}\right)^T w^{(k)}\right| \geq \alpha\right] \\
&\leq \sum_{j=1}^{p} 2\exp\left(-\frac{\alpha^2 n}{2\sigma^2\left(X_j^{(k)}\right)^T X_j^{(k)}}\right) \\
&\leq 2p\exp\left(-\frac{\alpha^2 n}{4\sigma^2}\right).
\end{aligned}$$

By union bound, the result follows.

□

2.6.2 Proof of Theorem 2

Let $d = \lfloor \frac{\lambda_b}{\lambda_s} \rfloor$ and $(B^*, S^*) = \mathcal{H}_d(\bar{\Theta})$. Then, the result follows from the next proposition.

Proposition 2. *Under assumptions of Theorem 2, if*

$$n > \max\left(\frac{Bs\log(pr)}{C_{min}\gamma_s^2}, \frac{Bsr\big(r\log(2)+\log(p)\big)}{C_{min}\gamma_b^2}\right)$$

then with probability at least $1 - c_1\exp\left(-c_2\left(r\log(2)+\log(p)\right)\right) - c_3\exp(-c_4\log(rs))$ *for some positive constants* $c_1 - c_4$*, we are guaranteed that the following properties hold:*

(P1) *The solution* $(\hat{B}, \hat{S})$ *to (2.1) is unique and* $RowSupp(\hat{B}) \subseteq RowSupp(B^*)$ *and* $Supp(\hat{S}) \subseteq Supp(S^*)$.

(P2) $$\left\|\hat{B} + \hat{S} - B^* - S^*\right\|_\infty \leq \underbrace{\sqrt{\frac{50\sigma^2 \log(rs)}{nC_{min}}} + \lambda_s \left(\frac{Ds}{C_{min}\sqrt{n}} + D_{max}\right)}_{g_{\min}}.$$

(P3) $sign\left(Supp(\hat{s}_j)\right) = sign\left(Supp(s_j^*)\right)$
for all $j \notin RowSupp(B^*)$ *provided that*

$$\min_{\substack{j \notin RowSupp(B^*) \\ (j,k) \in Supp(S^*)}} \left|s_j^{*(k)}\right| > g_{\min}.$$

(P4) $sign\Big(Supp(\hat{s}_j + \hat{b}_j)\Big) = sign\left(Supp(s_j^* + b_j^*)\right)$
for all $j \in RowSupp(B^*)$ *provided that*

$$\min_{(j,k) \in Supp(B^*)} \left|b_j^{*(k)} + s_j^{*(k)}\right| > g_{\min}.$$

Proof. We provide the proof of each part separately.

(P1) Considering the constructed primal-dual pair $(\tilde{S}, \tilde{B}, \widetilde{Z})$, it suffices to show that the conditions (C3) and (C4) in Lemma 33 are satisfied under these assumptions. Lemma 5 guarantees that with probability at least $1 - c_1 \exp\left(-c_2\left(r\log(2) + \log(p)\right)\right)$ those conditions are satisfied. Hence, $(\hat{B}, \hat{S}) = (\tilde{B}, \tilde{S})$ are the unique solution to (2.1) and (P1) follows.

(P2) From (2.5), we have

$$\begin{aligned}\max_{j\in\mathcal{U}_k}\left|\Delta_j^{(k)}\right| &\leq \underbrace{\left\|\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\frac{1}{n}\left(X_{\mathcal{U}_k}^{(k)}\right)^T w^{(k)}\right\|_\infty}_{\mathcal{W}^{(k)}} \\ &\quad + \left\|\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty \\ &\leq \left\|\mathcal{W}^{(k)}\right\|_\infty + \left\|\left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty \\ &\quad + \left\|\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1} - \left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right)\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty .\end{aligned}$$

We need to bound these three quantities. Notice that

$$\begin{aligned}\left\|\left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty &\leq \left\|\left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right\|_{\infty,1}\left\|\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty \\ &\leq D_{max}\lambda_s.\end{aligned}$$

Also, we have

$$\begin{aligned}&\left\|\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1} - \left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right)\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty \\ &\quad\leq \lambda_{max}\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1} - \left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right)\left\|\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_2 \\ &\quad\leq \lambda_{max}\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1} - \left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right)\sqrt{s}\lambda_s \\ &\quad\leq \frac{4}{C_{min}}\sqrt{\frac{s}{n}}\sqrt{s}\lambda_s,\end{aligned}$$

where, the last inequality holds with probability at least $1-c_1\exp\left(-c_2\left(\sqrt{n}-\sqrt{s}\right)^2\right)$ for some positive constants c_1 and c_2 as a result of [40] on eigenvalues of Gaussian random matrices. Conditioned on $X_{\mathcal{U}_k}^{(k)}$, the vector

$\mathcal{W}^{(k)} \in \mathbb{R}^{|\mathcal{U}_k|}$ is a zero-mean Gaussian random vector with covariance matrix $\frac{\sigma^2}{n}\left(\frac{1}{n}\left\langle X^{(k)}_{\mathcal{U}_k}, X^{(k)}_{\mathcal{U}_k}\right\rangle\right)^{-1}$. Thus, we have

$$\begin{aligned}
&\frac{1}{n}\lambda_{max}\left(\left(\frac{1}{n}\left\langle X^{(k)}_{\mathcal{U}_k}, X^{(k)}_{\mathcal{U}_k}\right\rangle\right)^{-1}\right) \\
&\quad\leq \frac{1}{n}\lambda_{max}\left(\left(\frac{1}{n}\left\langle X^{(k)}_{\mathcal{U}_k}, X^{(k)}_{\mathcal{U}_k}\right\rangle\right)^{-1} - \left(\Sigma^{(k)}_{\mathcal{U}_k,\mathcal{U}_k}\right)^{-1}\right) \\
&\qquad\qquad + \frac{1}{n}\lambda_{max}\left(\left(\Sigma^{(k)}_{\mathcal{U}_k,\mathcal{U}_k}\right)^{-1}\right) \\
&\quad\leq \frac{1}{n}\left(\frac{4}{C_{min}}\sqrt{\frac{s}{n}} + \frac{1}{C_{min}}\right) \\
&\quad\leq \frac{5}{nC_{min}}.
\end{aligned}$$

From the concentration of Gaussian random variables (Lemma 4) and using the union bound, we get

$$\mathbb{P}\left[\max_{1\leq k\leq r}\left\|\mathcal{W}^{(k)}\right\|_\infty \geq t\right] \leq 2\exp\left(-\frac{t^2 nC_{min}}{50\sigma^2} + \log(rs)\right).$$

For $t = \epsilon\sqrt{\frac{50\sigma^2\log(rs)}{nC_{min}}}$ for some $\epsilon > 1$, the result follows.

(P3),(P4) The results are immediate consequence of (P2).

□

Lemma 5. *Under the assumptions of Proposition 2, the conditions (C3) and (C4) in Lemma 33 hold for the constructed primal-dual pair with probability at least* $1 - c_1\exp\left(-c_2\left(r\log(2) + \log(p)\right)\right)$ *for some positive constants* c_1 *and* c_2.

Proof. First, we need to bound the projection of $\widetilde{Z}$ into the space U_s^c. Notice

that

$$\left|\left(P_{U_s^c}(\widetilde{Z})\right)_j^{(k)}\right| = \begin{cases} \dfrac{\lambda_b - \lambda_s\|\tilde{s}_j\|_0}{\left|M_j(\tilde{B})\right| - \|\tilde{s}_j\|_0} & \\ \qquad\qquad j \in \mathrm{RowSupp}(\tilde{B}) \quad \& \quad (j,k) \notin \mathrm{Supp}(\tilde{S}) & \\ \left|\tilde{z}_j^{(k)}\right| \qquad\qquad\qquad\qquad j \in \bigcap_{k=1}^r \mathcal{U}_k^c & \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad \text{ow.} \end{cases}.$$

By our assumptions on the ratio of the penalty regularizer coefficients, we have $\frac{\lambda_b - \lambda_s\|\tilde{s}_j\|_0}{\left|M_j(\tilde{B})\right| - \|\tilde{s}_j\|_0} < \lambda_s$. For all $j \in \bigcap_{k=1}^r \mathcal{U}_k$ and $R \in \mathbb{R}^{p\times r}$ with i.i.d. standard Gaussian entries (see Lemma 4 in [101]), we have

$$\begin{aligned}
&\left|\tilde{z}_j^{(k)}\right| \\
&\leq \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \underbrace{\left|\frac{1}{n}\left\langle X_j^{(k)}, \mathbf{I} - \frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\left(X_{\mathcal{U}_k}^{(k)}\right)^T\right\rangle w^{(k)}\right|}_{\mathcal{W}_j^{(k)}} \\
&\qquad + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle \tilde{z}_{\mathcal{U}_k}^{(k)}\right| \\
&\leq \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left\|\Sigma_{j,\mathcal{U}_k}^{(k)}\left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right\|_1 \left\|\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty \\
&\quad + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \underbrace{\left|\frac{1}{n}\left\langle R_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle \tilde{z}_{\mathcal{U}_k}^{(k)}\right|}_{\mathcal{R}_j^{(k)}} \\
&\leq (1-\gamma_s)\lambda_s + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{R}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right|,
\end{aligned}$$

The second inequality follows from the triangle inequality on the distributions. By Lemma 6, if $n \geq \frac{2}{2-\sqrt{3}}\log(pr)$ then with high probability $\left\|X_j^{(k)}\right\|_2^2 \leq 2n$ and hence $\mathrm{Var}\left(\mathcal{W}_j^{(k)}\right) \leq \frac{2\sigma^2}{n}$. Using the concentration results for the zero-mean

Gaussian random variable $\mathcal{W}_j^{(k)}$ and using the union bound, we get

$$\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right|\geq t\right]\leq 2\exp\left(-\frac{t^2n}{4\sigma^2}+\log(p)\right)\qquad\forall t\geq 0.$$

Conditioning on $\left(X_{\mathcal{U}_k}^{(k)},w^{(k)},\tilde{z}^{(k)}\right)$'s, we have that $\mathcal{R}_j^{(k)}$ is a zero-mean Gaussian random variable with

$$\mathrm{Var}\left(\mathcal{R}_j^{(k)}\right)\leq\frac{\left\|\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_2^2}{nC_{min}}\leq\frac{s\lambda_s^2}{nC_{min}}.$$

By concentration of Gaussian random variables, we have

$$\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{R}_j^{(k)}\right|\geq t\right]\leq 2\exp\left(-\frac{t^2nC_{min}}{Bs\lambda_s^2}+\log(p)\right)\quad\forall t\geq 0.$$

Using these bounds, we get

$$\begin{aligned}
&\mathbb{P}\left[\left\|P_{U_s^c}(\widetilde{Z})\right\|_{\infty,\infty}<\lambda_s\right]\\
&\geq\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{R}_j^{(k)}\right|+\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right|<\gamma_s\lambda_s\qquad\forall\,1\leq k\leq r\right]\\
&\geq\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{R}_j^{(k)}\right|<t_0\quad\forall\,1\leq k\leq r\right]\\
&\qquad\qquad\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right|<\gamma_s\lambda_s-t_0\quad\forall\,1\leq k\leq r\right]\\
&\geq\left(1-2\exp\left(-\frac{t_0^2nC_{min}}{Bs\lambda_s^2}+\log(pr)\right)\right)\\
&\qquad\qquad\left(1-2\exp\left(-\frac{(\gamma_s\lambda_s-t_0)^2n}{4\sigma^2}+\log(pr)\right)\right).
\end{aligned}$$

This probability goes to 1 for $t_0=\frac{\sqrt{Bs}\lambda_s}{\sqrt{Bs}\lambda_s+2\sigma\sqrt{C_{min}}}\gamma_s\lambda_s$ (the solution to $\frac{t_0^2C_{min}}{Bs\lambda_s^2}=\frac{(\gamma_s\lambda_s-t_0)^2}{4\sigma^2}$), if the regularization parameter $\lambda_s>\frac{\sqrt{4\sigma^2C_{min}\log(pr)}}{\gamma_s\sqrt{nC_{min}}-\sqrt{Bs\log(pr)}}$ provided that $n>\frac{Bs\log(pr)}{C_{min}\gamma_s^2}$ as stated in the assumptions.

Next, we need to bound the projection of $\widetilde{Z}$ into the space U_b^c. Notice that

$$\sum_{k=1}^{r}\left|\left(P_{U_b^c}(\widetilde{Z})\right)_j^{(k)}\right| = \begin{cases} \lambda_s\|\tilde{s}_j\|_0 & j \in \bigcup_{k=1}^{r}\mathcal{U}_k - \mathrm{RowSupp}(B^*) \\ \sum_{k=1}^{r}\left|\tilde{\tilde{z}}_j^{(k)}\right| & j \in \bigcap_{k=1}^{r}\mathcal{U}_k^c \\ 0 & \text{ow} \end{cases}.$$

We have $\lambda_s\|\tilde{s}_j\|_0 \leq \lambda_s D(S^*) < \lambda_b$ by our assumption on the ratio of the penalty regularizer coefficients. For all $j \in \bigcap_{k=1}^{r}\mathcal{U}_k^c$, we have

$$\begin{aligned}
&\sum_{k=1}^{r}\left|\tilde{\tilde{z}}_j^{(k)}\right| \\
&\leq \max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\underbrace{\left|\frac{1}{n}\left\langle X_j^{(k)}, \mathbf{I}-\frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\left(X_{\mathcal{U}_k}^{(k)}\right)^T\right\rangle w^{(k)}\right|}_{\mathcal{W}_j^{(k)}} \\
&\quad + \max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle \tilde{z}_{\mathcal{U}_k}^{(k)}\right| \\
&\leq \max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{W}_j^{(k)}\right| \\
&\quad + \max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left\|\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle\right\|_1 \\
&\qquad\qquad \max_{j\in\bigcup_{k=1}^{r}\mathcal{U}_k}\left\|\tilde{\tilde{z}}_j^{(k)}\right\|_1 \\
&\quad + \max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\underbrace{\left|\frac{1}{n}\left\langle R_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle \tilde{z}_{\mathcal{U}_k}^{(k)}\right|}_{\mathcal{R}_j^{(k)}} \\
&\leq (1-\gamma_b)\lambda_b + \max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{R}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{W}_j^{(k)}\right|.
\end{aligned}$$

Let $\mathbf{v} \in \{-1,+1\}^r$ be a vector of signs such that $\sum_{k=1}^{r}\left|\mathcal{W}_j^{(k)}\right| = \sum_{k=1}^{r} v_k\mathcal{W}_j^{(k)}$.

Then,

$$\mathrm{Var}\left(\sum_{k=1}^{r}\left|\mathcal{W}_j^{(k)}\right|\right) = \mathrm{Var}\left(\sum_{k=1}^{r} v_k \mathcal{W}_j^{(k)}\right) \leq \frac{2\sigma^2 r}{n}.$$

Using the union bound and previous discussion, we get

$$\begin{aligned}
&\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{W}_j^{(k)}\right| \geq t\right] \\
&\quad = \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\max_{\mathbf{v}\in\{-1,+1\}^r}\sum_{k=1}^{r} v_k \mathcal{W}_j^{(k)} \geq t\right] \\
&\quad \leq 2\exp\left(-\frac{t^2 n}{4\sigma^2 r} + r\log(2) + \log(p)\right) \qquad \forall t \geq 0.
\end{aligned}$$

We have

$$\begin{aligned}
\mathrm{Var}\left(\sum_{k=1}^{r}\left|\mathcal{R}_j^{(k)}\right|\right) &= \mathrm{Var}\left(\sum_{k=1}^{r} v_k \mathcal{R}_j^{(k)}\right) \\
&\leq \frac{\sum_{k=1}^{r}\left\|\tilde{z}_j^{(k)}\right\|_2^2}{nC_{min}} \leq \frac{rs\lambda_s^2}{nC_{min}} < \frac{rs\lambda_b^2}{nC_{min}}
\end{aligned}$$

and consequently by concentration of Gaussian variables,

$$\begin{aligned}
&\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{K}\left|\mathcal{R}_j^{(k)}\right| \geq t\right] \\
&\quad = \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\max_{\mathbf{v}\in\{-1,+1\}^r}\sum_{k=1}^{r} v_k \mathcal{R}_j^{(k)} \geq t\right] \\
&\quad \leq 2\exp\left(-\frac{t^2 nC_{min}}{2rs\lambda_b^2} + r\log(2) + \log(p)\right) \qquad \forall t \geq 0.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\mathbb{P}\left[\left\|P_{U_b^c}(\widetilde{Z})\right\|_{\infty,1} < \lambda_b\right] \\
&\quad \geq \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \sum_{k=1}^r \left|\mathcal{R}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \sum_{k=1}^r \left|\mathcal{W}_j^{(k)}\right| < \gamma_b\lambda_b\right] \\
&\quad \geq \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \sum_{k=1}^r \left|\mathcal{R}_j^{(k)}\right| < t_0\right] \\
&\qquad\qquad \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \sum_{k=1}^r \left|\mathcal{W}_j^{(k)}\right| < \gamma_b\lambda_b - t_0\right] \\
&\quad \geq \left(1 - 2\exp\left(-\frac{t_0^2 nC_{min}}{2rs\lambda_b^2} + r\log(2) + \log(p)\right)\right) \\
&\qquad\qquad \left(1 - 2\exp\left(-\frac{(\gamma_b\lambda_b - t_0)^2 n}{4\sigma^2 r} + r\log(2) + \log(p)\right)\right).
\end{aligned}$$

This probability goes to 1 for $t_0 = \frac{\sqrt{Bs}\lambda_b}{\sqrt{Bs}\lambda_b + 2\sigma\sqrt{C_{min}}}\gamma_b\lambda_b$ (the solution to $\frac{(\gamma_b\lambda_b - t_0)^2 n}{4\sigma^2 r} = \frac{t_0^2 nC_{min}}{2rs\lambda_b^2}$), if

$$\lambda_b > \frac{\sqrt{4\sigma^2 C_{min} r\Big(r\log(2) + \log(p)\Big)}}{\gamma_b\sqrt{nC_{min}} - \sqrt{Bsr\Big(r\log(2) + \log(p)\Big)}},$$

provided that $n > \frac{Bsr(r\log(2)+\log(p))}{\gamma_b^2 C_{min}}$ as stated in the assumptions. Hence, with probability at least $1 - c_1\exp\left(-c_2\left(r\log(2) + \log(p)\right)\right)$ the conditions of the Lemma 33 are satisfied.

□

Lemma 6.

$$\mathbb{P}\left[\max_{1\leq k\leq r}\max_{1\leq j\leq p}\left\|X_j^{(k)}\right\|_2^2 \leq 2n\right] \geq 1 - \exp\left(-(1-\frac{\sqrt{3}}{2})n + \log(pr)\right).$$

Proof. Notice that $\|X_j^{(k)}\|_2^2$ is a χ^2 random variable with n degrees of freedom. According to [80], we have

$$\mathbb{P}\left[\left\|X_j^{(k)}\right\|_2^2 \geq t + (\sqrt{t}+\sqrt{n})^2\right] \leq \exp(-t) \qquad \forall t \geq 0.$$

Letting $t = \left(\frac{\sqrt{3}-1}{2}\right)^2 n$ and using the union bound, the result follows. □

2.6.3 Proof of Theorem 3

We will actually prove a more general theorem, from which Theorem 3 would follow as a corollary. Among shared features (with size αs), we say a fraction τ has different magnitudes on $\bar{\Theta}$. Let τ_1 be the fraction with larger magnitude on the first task and τ_2 the fraction with larger magnitude on the second task (so that $\tau = \tau_1 + \tau_2$). Moreover, let $\frac{\lambda_b}{\lambda_s} = \kappa$ and

$$f(\kappa) = f(\kappa, \tau, \alpha) = 2 - 2(1-\tau)\alpha - 2\tau\alpha\kappa + \left(\frac{1+\tau}{2}\right)\alpha\kappa^2,$$

and

$$g(\kappa, \tau, \alpha) = \max\left(\frac{2\, f(\kappa)}{\kappa^2}, f(\kappa)\right).$$

Theorem 4. *Under the assumptions of the Theorem 3, if*

$$\left|\left\{j \in \mathit{RowSupp}(B^*) : \left|\left|\Theta_j^{*(1)}\right| - \left|\Theta_j^{*(2)}\right|\right| \leq c\lambda_s\right\}\right| = (1-\tau)\alpha s,$$

then, the result of Theorem 3 holds for

$$\theta(n, s, p, \alpha) = \frac{n}{g(\kappa, \tau, \alpha)\, s \log\left(p - (2-\alpha)s\right)}.$$

Corollary 4. *Under the assumptions of the Theorem 4, if the regularization penalties are set as $\kappa = \lambda_b/\lambda_s = \sqrt{2}$, then the result of Theorem 3 holds for $\theta(n, s, p, \alpha) = \frac{n}{\left(2-\alpha+(3-2\sqrt{2})\tau\alpha\right) s \log(p-(2-\alpha)s)}$.*

Proof. Follows trivially by substituting $\kappa = \sqrt{2}$ in Theorem 4. Indeed, this setting of κ can also be shown to minimize $g(\kappa, \tau, \alpha)$:

$$\begin{aligned}
&\min_{1<\kappa<2} \max\left(\frac{2\,f(\kappa)}{\kappa^2}, f(\kappa)\right) \\
&= \min\left(\min_{1<\kappa\leq\sqrt{2}} \frac{2}{\kappa^2}\left(f(\kappa)\right), \min_{\sqrt{2}<\kappa<2} f(\kappa)\right) \\
&= 2 - \alpha + (3 - 2\sqrt{2})\,\tau\,\alpha.
\end{aligned}$$

□

Proof of Theorem 3: The proof follows from Corollary 4 by setting $\tau = 0$ and $\kappa = \sqrt{2}$.

We will now set out to prove Theorem 4. We will first need the following lemma.

Lemma 7. *For any $j \in RowSupp(B^*)$, if $\left|S_j^{*(k)}\right| < c\lambda_s$ for some constant c specified in the proof, then $\tilde{S}_j^{(k)} = 0$ with probability $1 - c_1 \exp(-c_2 n)$.*

Proof. Let $\check{S}$ be a matrix equal to $\tilde{S}$ except that $\check{S}_j^{(k)} = 0$. Using the concen-

tration of Gaussian random variables and optimality of $\tilde{S}$, we get

$$\begin{aligned}
&\mathbb{P}\left[\left|\tilde{S}_j^{(k)}\right| > 0\right] \\
&\leq \mathbb{P}\Bigg[2n\lambda_s \left|\tilde{S}_j^{(k)}\right| < \left\|y^{(k)} - X^{(k)}(\tilde{B}^{(k)} + \check{S}^{(k)})\right\|_2^2 \\
&\qquad\qquad - \left\|y^{(k)} - X^{(k)}(\tilde{B}^{(k)} + \tilde{S}^{(k)})\right\|_2^2\Bigg] \\
&= \mathbb{P}\Bigg[2n\lambda_s < \Bigg(\frac{\left\|y^{(k)} - X^{(k)}(\tilde{B}^{(k)} + \check{S}^{(k)})\right\|_2^2}{\left\|\tilde{S}_j^{(k)} X_j^{(k)}\right\|_2} \\
&\qquad - \frac{\left\|y^{(k)} - X^{(k)}(\tilde{B}^{(k)} + \check{S}^{(k)}) - \tilde{S}_j^{(k)} X_j^{(k)}\right\|_2^2}{\left\|\tilde{S}_j^{(k)} X_j^{(k)}\right\|_2}\Bigg) \left\|X_j^{(k)}\right\|_2\Bigg] \\
&\leq \mathbb{P}\left[2n\lambda_s < 2\left\|X_j^{(k)}\right\|_2^2 \left\|y^{(k)} - X^{(k)}(\tilde{B}^{(k)} + \check{S}^{(k)})\right\|_2\right] \\
&= \mathbb{P}\left[n\lambda_s < \left\|X_j^{(k)}\right\|_2^2 \left\|X^{(k)}(B^{*(k)} + S^{*(k)} - \tilde{B}^{(k)} - \check{S}^{(k)}) + w^{(k)}\right\|_2\right]
\end{aligned}$$

Using the ℓ_∞ bound on the error, for some constant c, we have

$$\begin{aligned}
\mathbb{P}\left[\left|\tilde{S}_j^{(k)}\right| > 0\right] &\leq \mathbb{P}\left[n\lambda_s < \frac{1}{c}\left|S_j^{*(k)}\right| \left\|X_j^{(k)}\right\|_2^2\right] \\
&= \mathbb{P}\left[\frac{c\lambda_s}{\left|S_j^{*(k)}\right|} n < \left\|X_j^{(k)}\right\|_2^2\right].
\end{aligned}$$

Notice that $\mathbb{E}[\|X_j^{(k)}\|_2^2] = n$. According to the concentration of χ^2 random variables concentration theorems (see [80]), this probability vanishes exponentially fast in n for $\left|\bar{S}_j^{(k)}\right| < c\lambda_s$.

□

2.6.4 Proof of Theorem 4

We will now provide the proofs of different parts separately.

Proof. **(Success):** Recall the constructed primal-dual pair $(\tilde{B}, \tilde{S}, \widetilde{Z})$. It suffices to show that the dual variable $\widetilde{Z}$ satisfies the conditions (C3) and (C4) of Lemma 33. By Lemma 8, these conditions are satisfied with probability at least $1 - c_1 \exp(-c_2 n)$ for some positive constants c_1 and c_2. Hence, $(\hat{B}, \hat{S}) = (\tilde{B}, \tilde{S})$ is the unique optimal solution. The rest are direct consequences of Proposition 2 for $C_{min} = 1$ and $D_{max} = 1$.

(Failure): We prove this result by contradiction. Suppose there exist a solution to (2.1), say $(\hat{B}, \hat{S})$ such that $\text{sign}\left(\text{Supp}(\hat{B} + \hat{S})\right) = \text{sign}\left(\text{Supp}(B^* + S^*)\right)$. By Lemma 11, this is equivalent to having $\text{sign}\left(\text{Supp}(\hat{B})\right) = \text{sign}\left(\text{Supp}(B^*)\right)$ and $\text{sign}\left(\text{Supp}(\hat{S})\right) = \text{sign}\left(\text{Supp}(S^*)\right)$ and $\frac{\lambda_b}{\lambda_s} = \kappa$.

Now, suppose $n < (1-\nu) \max\left(\frac{2\, f(\kappa)}{\kappa^2}, f(\kappa)\right) s \log(p - (2-\alpha)s)$, for some $\nu > 0$. This entails that

either (i) $n < (1-\nu) f(\kappa) s \log(p - (2-\alpha)s)$,

or (ii) $n < (1-\nu) \left(\frac{2\, f(\kappa)}{\kappa^2}\right) s \log(p - (2-\alpha)s)$.

Case (i): We will show that with high probability, there exists k for which, there exists $j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c$ such that $\left|\tilde{Z}_j^{(k)}\right| > \lambda_s$. This is a contradiction to Lemma 36.

Using (2.6) and conditioning on $(X_{\mathcal{U}_k}^{(k)}, w^{(k)}, \tilde{Z}_{\mathcal{U}_k}^{(k)})$, for all $j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c$ we have that the random variables $\tilde{Z}_j^{(k)}$ are i.i.d. zero-mean Gaussian random

variables with

$$\begin{aligned}
&\mathrm{Var}\left(\tilde{Z}_j^{(k)}\right) \\
&= \Bigg\| \frac{1}{n} X_{\mathcal{U}_k}^{(k)} \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \tilde{Z}_{\mathcal{U}_k}^{(k)} \\
&\qquad + \frac{1}{n} \left(\mathbf{I} - \frac{1}{n} X_{\mathcal{U}_k}^{(k)} \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \left(X_{\mathcal{U}_k}^{(k)} \right)^T \right) w^{(k)} \Bigg\|_2^2 \\
&= \left\| \frac{1}{n} X_{\mathcal{U}_k}^{(k)} \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \tilde{Z}_{\mathcal{U}_k}^{(k)} \right\|_2^2 \\
&\qquad + \left\| \frac{1}{n} \left(\mathbf{I} - \frac{1}{n} X_{\mathcal{U}_k}^{(k)} \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \left(X_{\mathcal{U}_k}^{(k)} \right)^T \right) w^{(k)} \right\|_2^2
\end{aligned}$$

The second equality holds by orthogonality of projections. We thus have

$$\begin{aligned}
&\mathrm{Var}\left(\tilde{Z}_j^{(k)}\right) \\
&\geq \max \Bigg(\lambda_{min} \left(\left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \right) \frac{\left\| \tilde{Z}_{\mathcal{U}_k}^{(k)} \right\|_2^2}{n} \\
&\qquad , \frac{\left\| \left(\mathbf{I} - \frac{1}{n} X_{\mathcal{U}_k}^{(k)} \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \left(X_{\mathcal{U}_k}^{(k)} \right)^T \right) w^{(k)} \right\|_2^2}{n^2} \Bigg) \\
&\geq \frac{\left\| \tilde{Z}_{\mathcal{U}_k}^{(k)} \right\|_2^2}{\left(\sqrt{n} + \sqrt{s} \right)^2}
\end{aligned}$$

The second inequality holds with probability at least $1 - c_1 \exp\left(-c_2 \left(\sqrt{n} + \sqrt{s}\right)^2\right)$ as a result of [40] on the eigenvalues of Gaussian matrices. The third inequality holds with probability at least $1 - c_3 \exp(-c_4 n)$ as a result of [80] on the magnitude of χ^2 random variables. Considering $\tilde{B} + \tilde{S}$, assume that among shared features (with size αs), a portion of τ_1 has larger magnitude on the fist task and a portion of τ_2 has larger magnitude on the second task (and

consequently a portion of $1 - \tau_1 - \tau_2$ has equal magnitude on both tasks). Assuming $\lambda_b = \kappa \lambda_s$ for some $\kappa \in (1, 2)$, we get

$$\begin{aligned}
\widetilde{\sigma}_1^2 &:= \mathrm{Var}\left(\tilde{Z}_j^{(1)}\right) \\
&= \frac{(1-\alpha)s\lambda_s^2 + \tau_1 \alpha s \lambda_s^2 + \tau_2 \alpha s (\lambda_b - \lambda_s)^2 + (1 - \tau_1 - \tau_2)\alpha s \frac{\lambda_b^2}{4}}{(\sqrt{n} + \sqrt{s})^2} \\
&=: \frac{f_1(\kappa) s \lambda_s^2}{n\,\left(1 + \sqrt{\frac{s}{n}}\right)^2}.
\end{aligned}$$

The first equality follows from the construction of the dual matrix and the fact that we have recovered the sign support correctly. The last strict inequality follows from the assumption that $\theta(n, p, s, \alpha) < 1$. Similarly, we have

$$\begin{aligned}
\widetilde{\sigma}_2^2 &:= \mathrm{Var}\left(\tilde{Z}_j^{(2)}\right) \\
&> \frac{(1-\alpha)s\lambda_s^2 + \tau_2 \alpha s \lambda_s^2 + \tau_1 \alpha s (\lambda_b - \lambda_s)^2 + (1 - \tau_1 - \tau_2)\alpha s \frac{\lambda_b^2}{4}}{n\,\left(1 + \sqrt{\frac{s}{n}}\right)^2} \\
&=: \frac{f_2(\kappa) s \lambda_s^2}{n\,\left(1 + \sqrt{\frac{s}{n}}\right)^2}.
\end{aligned}$$

Given these lower bounds on the variance, by results on Gaussian maxima (see [40]), for any $\delta > 0$, with high probability,

$$\begin{aligned}
&\max_{1 \le k \le r} \max_{j \in \bigcup_{k=1}^r \mathcal{U}_k} \left|\tilde{Z}_j^{(k)}\right| \\
&\qquad \ge (1 - \delta)\sqrt{(\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2) \log\left(r\Big(p - (2-\alpha)s\Big)\right)}.
\end{aligned}$$

This in turn can be bound as

$$\begin{aligned}
&(1-\delta)\,(\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2) \log\left(r\Big(p - (2-\alpha)s\Big)\right) \\
&\qquad \ge (1-\delta)\frac{(f_1(\kappa) + f_2(\kappa))\ s\ \log\left(r\Big(p - (2-\alpha)s\Big)\right)}{n\,\left(1 + \sqrt{\frac{s}{n}}\right)^2}\lambda_s^2. \\
&\qquad \ge (1-\delta)\frac{f(\kappa)\, s\ \log\left(r\Big(p - (2-\alpha)s\Big)\right)}{n\,\left(1 + \sqrt{\frac{s}{n}}\right)^2}\lambda_s^2.
\end{aligned}$$

Consider two cases:

1. $\frac{s}{n} = \Omega(1)$: In this case, we have $s > cn$ for some constant $c > 0$. Then,

$$\begin{aligned}
&(1-\delta)\frac{(f(\kappa))\ s\ \log\Big(r\Big(p-(2-\alpha)s\Big)\Big)}{n\ \left(1+\sqrt{\frac{s}{n}}\right)^2}\lambda_s^2 \\
&\qquad = (1-\delta)\frac{(f(\kappa))\ (s/n)\ \log\Big(r\Big(p-(2-\alpha)s\Big)\Big)}{\left(1+\sqrt{s/n}\right)^2}\lambda_s^2 \\
&\qquad > c' f(\kappa)\ \log\Big(r\Big(p-(2-\alpha)s\Big)\Big)\ \lambda_s^2 \\
&\qquad > (1+\epsilon)\lambda_s^2,
\end{aligned}$$

   for any fixed $\epsilon > 0$, as $p \to \infty$.

2. $\frac{s}{n} \to 0$: In this case, we have $s/n = o(1)$. Here we will use that the sample size scales as $n < (1-\nu)\,(f(\kappa))\, s \log(p-(2-\alpha)s)$.

$$\begin{aligned}
&(1-\delta)\frac{(f(\kappa))\ s\ \log\Big(r\Big(p-(2-\alpha)s\Big)\Big)}{n\ \left(1+\sqrt{\frac{s}{n}}\right)^2}\lambda_s^2 \\
&\qquad \geq \frac{(1-\delta)(1-o(1))}{1-\nu}\lambda_s^2 \\
&\qquad > (1+\epsilon)\lambda_s^2,
\end{aligned}$$

   for some $\epsilon > 0$ by taking δ small enough.

Thus with high probability, $\exists k \exists j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c$ such that $\left|\tilde{Z}_j^{(k)}\right| > \lambda_s$. This is a contradiction to Lemma 36.

Case (ii): We need to show that with high probability, there exist a row that violates the sub-gradient condition of ℓ_∞-norm: $\exists j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c$ such

that $\left\|\tilde{Z}_j^{(k)}\right\|_1 > \lambda_b$. This is a contradiction to Lemma 36.

Following the same proof technique, notice that $\sum_{k=1}^r \tilde{Z}_j^{(k)}$ is a zero-mean Gaussian random variable with $\mathrm{Var}\left(\sum_{k=1}^r \tilde{Z}_j^{(k)}\right) \geq r(\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2)$. Thus, with high probability

$$\max_{j \in \bigcap_{k=1}^r \mathcal{U}_k^c} \left\|\tilde{Z}_j^{(k)}\right\|_1 \geq (1-\delta)\sqrt{r(\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2)\log\Big(p - (2-\alpha)s\Big)}.$$

Following the same line of argument for this case, yields the required bound $\left\|\tilde{Z}_j^{(k)}\right\|_1 > (1+\epsilon)\lambda_b$.

This concludes the proof of the theorem.

□

Lemma 8. *Under assumptions of Theorem 3, the conditions (C3) and (C4) in Lemma 33 hold with probability at least* $1 - c_1\exp(-c_2 n)$ *for some positive constants* c_1 *and* c_2.

Proof. First, we need to bound the projection of $\widetilde{Z}$ into the space U_s^c. Notice that

$$\left|\Big(P_{U_s^c}(\widetilde{Z})\Big)_j^{(k)}\right| = \begin{cases} \dfrac{\lambda_b - \lambda_s\|\tilde{S}_j\|_0}{\left|M_j(\tilde{B})\right| - \|\tilde{S}_j\|_0} & \\ \qquad j \in \mathrm{RowSupp}(\tilde{B}) \quad \& \quad (j,k) \notin \mathrm{Supp}(\tilde{S}) & \\ \left|\tilde{Z}_j^{(k)}\right| \qquad\qquad j \in \bigcap_{k=1}^r \mathcal{U}_k^c & \\ 0 \qquad\qquad\qquad\qquad \text{ow.} \end{cases}.$$

By our assumption on the penalty regularizer coefficients, we have $\frac{\lambda_b - \lambda_s\|\tilde{S}_j\|_0}{\left|M_j^{\pm}(\tilde{B})\right| - \|\tilde{S}_j\|_0} <$

λ_s. Moreover, we have

$$\begin{aligned}
&\left|\tilde{Z}_j^{(k)}\right| \\
&\leq \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \underbrace{\left|\frac{1}{n}\left\langle X_j^{(k)}, \mathbf{I}-\frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\left(X_{\mathcal{U}_k}^{(k)}\right)^T\right\rangle w^{(k)}\right|}_{\mathcal{W}_j^{(k)}} \\
&\quad + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \underbrace{\left|\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle \tilde{Z}_{\mathcal{U}_k}^{(k)}\right|}_{\mathcal{Z}_j^{(k)}} \\
&\triangleq \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{Z}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right|.
\end{aligned}$$

By Lemma 6, if $n \geq \frac{2}{2-\sqrt{3}}\log(pK)$ then with high probability $\left\|X_j^{(k)}\right\|_2^2 \leq 2n$ and hence $\mathrm{Var}\left(\mathcal{W}_j^{(k)}\right) \leq \frac{2\sigma^2}{n}$. Notice that $\mathbb{E}\left[\left\|X_j^{(k)}\right\|_2^2\right] = n$ and we added the factor of 2 arbitrarily to use the concentration theorems. Using the concentration results for the zero-mean Gaussian random variable $\mathcal{W}_j^{(k)}$ and using the union bound, for all $t > 0$, we get

$$\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right| \geq t\right] \leq 2\exp\left(-\frac{t^2 n}{4\sigma^2} + \log\left(p-(2-\alpha)s\right)\right).$$

Conditioning on $\left(X_{\mathcal{U}_k}^{(k)}, w^{(k)}, \tilde{Z}^{(k)}\right)$'s, we have that $\mathcal{Z}_j^{(k)}$ is a zero-mean Gaussian random variable with

$$\mathrm{Var}\left(\mathcal{Z}_j^{(k)}\right) \leq \frac{1}{n}\lambda_{max}\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right)\left\|\tilde{Z}_{\mathcal{U}_k}^{(k)}\right\|_2^2.$$

According to the result of [40] on singular values of Gaussian matrices, for the matrix $X_{\mathcal{U}_k}^{(k)}$, for all $\delta > 0$, we have

$$\mathbb{P}\left[\sigma_{min}\left(X_{\mathcal{U}_k}^{(k)}\right) \leq (1-\delta)\left(\sqrt{n}-\sqrt{s}\right)\right] \leq \exp\left(-\frac{\delta^2\left(\sqrt{n}-\sqrt{s}\right)^2}{2}\right),$$

and since $\lambda_{max}\left(\left(\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right) = \sigma_{min}\left(X_{\mathcal{U}_k}^{(k)}\right)^{-2}$, we get

$$\mathbb{P}\left[\lambda_{max}\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right) \geq \frac{(1+\delta)}{\left(1-\sqrt{\frac{s}{n}}\right)^2}\right] \leq \exp\left(-\frac{\left(\sqrt{\delta+1}-1\right)^2\left(\sqrt{n}-\sqrt{s}\right)^2}{2(1+\delta)}\right).$$

According to Lemma 7, if $\left|\left|\Theta_j^{*(1)}\right| - \left|\Theta_j^{*(2)}\right|\right| = o(\lambda_s)$, then with high probability $\tilde{S}_j = 0$, so that $|\tilde{\Theta}_j^{(1)}| = |\tilde{\Theta}_j^{(2)}|$. Thus, among shared features (with size αs), a fraction τ have differing magnitudes on $\tilde{\Theta}$. Let τ_1 be the fraction with larger magnitude on the first task and τ_2 the fraction with larger magnitude on the second task (so that $\tau = \tau_1 + \tau_2$). Then, with high probability, recalling that $\lambda_b = \kappa\lambda_s$ for some $1 < \kappa < 2$, we get

$$\begin{aligned}
\mathrm{Var}\left(\mathcal{Z}_j^{(1)}\right) &\leq \frac{\left\|\tilde{Z}_{\mathcal{U}_1}^{(1)}\right\|_2^2}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&= \frac{(1-\alpha)s\lambda_s^2 + \tau_1\alpha s\lambda_s^2 + \tau_2\alpha s(\lambda_b-\lambda_s)^2 + (1-\tau_1-\tau_2)\alpha s\frac{\lambda_b^2}{4}}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&= \frac{\left(1-(1-\tau_1-\tau_2)\alpha - 2\tau_2\alpha\kappa + \left(\tau_2 + \frac{1-\tau_1-\tau_2}{4}\right)\alpha\kappa^2\right)s\lambda_s^2}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&\triangleq \frac{f_1(\kappa)s\lambda_s^2}{\left(\sqrt{n}-\sqrt{s}\right)^2}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{Var}\left(\mathcal{Z}_j^{(2)}\right) &\leq \frac{\left\|\tilde{Z}_{\mathcal{U}_2}^{(2)}\right\|_2^2}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&= \frac{\left(1-(1-\tau_1-\tau_2)\alpha - 2\tau_1\alpha\kappa + \left(\tau_1 + \frac{1-\tau_1-\tau_2}{4}\right)\alpha\kappa^2\right)s\lambda_s^2}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&\triangleq \frac{f_2(\kappa)s\lambda_s^2}{\left(\sqrt{n}-\sqrt{s}\right)^2}.
\end{aligned}$$

By concentration of Gaussian random variables, we have

$$\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{Z}_j^{(k)}\right|\geq t\right]$$
$$\leq 2\exp\left(-\frac{t^2\left(\sqrt{n}-\sqrt{s}\right)^2}{2f_k(\kappa)s\lambda_s^2}+\log\left(p-(1-\alpha)s\right)\right)\qquad\forall t\geq 0.$$

Using these bounds, we get

$$\begin{aligned}
&\mathbb{P}\left[\left\|P_{U_s^c}(\widetilde{Z})\right\|_{\infty,\infty}<\lambda_s\right]\\
&\geq\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{Z}_j^{(k)}\right|+\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right|<\lambda_s\qquad\forall\,1\leq k\leq K\right]\\
&\geq\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{Z}_j^{(k)}\right|<t_0\quad\forall\,1\leq k\leq r\right]\\
&\qquad\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right|<\lambda_s-t_0\quad\forall\,1\leq k\leq r\right]\\
&\geq\left(1-2\exp\left(-\frac{t_0^2\left(\sqrt{n}-\sqrt{s}\right)^2}{\left(f_1(\kappa)+f_2(\kappa)\right)s\lambda_s^2}+\log\left(p-(2-\alpha)s\right)+\log(r)\right)\right)\\
&\qquad\left(1-2\exp\left(-\frac{(\lambda_s-t_0)^2n}{4\sigma^2}+\log\left(p-(2-\alpha)s\right)+\log(r)\right)\right).
\end{aligned}$$

This probability goes to 1 for

$$t_0=\frac{\sqrt{\left(f_1(\kappa)+f_2(\kappa)\right)ns}\lambda_s}{\sqrt{\left(f_1(\kappa)+f_2(\kappa)\right)ns}\lambda_s+2\sigma(\sqrt{n}-\sqrt{s})}\lambda_s$$

(the solution to $\frac{t_0^2\left(\sqrt{n}-\sqrt{s}\right)^2}{(f_1(\kappa)+f_2(\kappa))s\lambda_s^2}=\frac{(\lambda_s-t_0)^2n}{4\sigma^2}$), if

$$\lambda_s>\frac{\sqrt{4\sigma^2\left(1-\sqrt{\frac{s}{n}}\right)^2\Big(\log(r)+\log\left(p-(2-\alpha)s\right)\Big)}}{\sqrt{n}-\left(\sqrt{s}+\sqrt{\left(f_1(\kappa)+f_2(\kappa)\right)s\Big(\log(r)+\log\left(p-(2-\alpha)s\right)\Big)}\right)}$$

provided that (substituting $r = 2$),

$$n > (f_1(\kappa) + f_2(\kappa))\, s \log\Big(p - (2-\alpha)s\Big) + \Bigg(1 + (f_1(\kappa) + f_2(\kappa)) \log(2) + 2\sqrt{(f_1(\kappa) + f_2(\kappa))\left(\log(2) + \log\Big(p - (2-\alpha)s\Big)\right)}\Bigg) s.$$

Since $f_1(\kappa) + f_2(\kappa) = f(\kappa)$ by definition, for large enough p with $\frac{s}{p} = \mathbf{o}(1)$, we require

$$n > f(\kappa) s \log\Big(p - (2-\alpha)s\Big). \tag{2.7}$$

Next, we need to bound the projection of $\widetilde{Z}$ into the space U_b^c. Notice that

$$\sum_{k=1}^{r} \left| \left(P_{U_b^c}(\widetilde{Z}) \right)_j^{(k)} \right| = \begin{cases} \lambda_s \|\tilde{S}_j\|_0 & j \in \bigcup_{k=1}^{r} \mathcal{U}_k - \mathrm{RowSupp}(B^*) \\ \sum_{k=1}^{r} \left| \tilde{Z}_j^{(k)} \right| & j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c \\ 0 & \text{ow} \end{cases}.$$

We have $\lambda_s \|\tilde{S}_j\|_0 \leq \lambda_s D(S^*) < \lambda_b$ by our assumption on the ratio of penalty

regularizer coefficients. For all $j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c$, we have

$$\begin{aligned}
&\sum_{k=1}^{r} \left| \tilde{Z}_j^{(k)} \right| \\
&\underset{j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c}{\leq \max} \sum_{k=1}^{r} \underbrace{\left| \frac{1}{n} \left\langle X_j^{(k)}, \mathbf{I} - \frac{1}{n} X_{\mathcal{U}_k}^{(k)} \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \left(X_{\mathcal{U}_k}^{(k)} \right)^T \right\rangle w^{(k)} \right|}_{\mathcal{W}_j^{(k)}} \\
&\qquad + \max_{j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c} \sum_{k=1}^{r} \underbrace{\left| \frac{1}{n} \left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)} \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \right\rangle \tilde{Z}_{\mathcal{U}_k}^{(k)} \right|}_{\mathcal{Z}_j^{(k)}} \\
&= \max_{j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c} \sum_{k=1}^{r} \left| \mathcal{Z}_j^{(k)} \right| + \max_{j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c} \sum_{k=1}^{r} \left| \mathcal{W}_j^{(k)} \right|.
\end{aligned}$$

Let $\mathbf{v} \in \{-1, +1\}^r$ be a vector of signs such that $\sum_{k=1}^{r} \left| \mathcal{W}_j^{(k)} \right| = \sum_{k=1}^{r} v_k \mathcal{W}_j^{(k)}$. Thus,

$$\operatorname{Var}\left(\sum_{k=1}^{r} \left| \mathcal{W}_j^{(k)} \right| \right) = \operatorname{Var}\left(\sum_{k=1}^{r} v_k \mathcal{W}_j^{(k)} \right) \leq \frac{2\sigma^2 r}{n}.$$

Using the union bound and previous discussion, for all $t > 0$, we get

$$\begin{aligned}
&\mathbb{P}\left[\max_{j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c} \sum_{k=1}^{r} \left| \mathcal{W}_j^{(k)} \right| \geq t \right] \\
&\qquad = \mathbb{P}\left[\max_{j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c} \max_{\mathbf{v} \in \{-1,+1\}^r} \sum_{k=1}^{r} v_k \mathcal{W}_j^{(k)} \geq t \right] \\
&\qquad \leq 2 \exp\left(-\frac{t^2 n}{4\sigma^2 r} + r \log(2) + \log\left(p - (2 - \alpha) s \right) \right).
\end{aligned}$$

Also from the previous analysis, assuming $\lambda_b = \kappa \lambda_s$ for some $1 < \kappa < 2$, we

get

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{k=1}^{r}\left|\mathcal{Z}_j^{(k)}\right|\right) = \mathrm{Var}\left(\sum_{k=1}^{r} v_k \mathcal{Z}_j^{(k)}\right) \leq \frac{\sum_{k=1}^{r}\left\|\tilde{Z}_j^{(k)}\right\|_2^2}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&= \frac{2(1-\alpha)s\lambda_s^2 + (\tau_1+\tau_2)\alpha s\lambda_s^2 + (\tau_1+\tau_2)\alpha s(\lambda_b-\lambda_s)^2 + 2(1-\tau_1-\tau_2)\alpha s\frac{\lambda_b^2}{4}}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&= \frac{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)s\lambda_b^2}{\left(\sqrt{n}-\sqrt{s}\right)^2}.
\end{aligned}$$

and consequently for all $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{Z}_j^{(k)}\right| \geq t\right] \\
&= \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\max_{\mathbf{v}\in\{-1,+1\}^r}\sum_{k=1}^{r} v_k\mathcal{Z}_j^{(k)} \geq t\right] \\
&\leq 2\exp\left(-\frac{t^2\left(\sqrt{n}-\sqrt{s}\right)^2}{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)s\lambda_b^2} + r\log(2) + \log\left(p-(2-\alpha)s\right)\right).
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\mathbb{P}\left[\left\|P_{U_b^c}(\widetilde{Z})\right\|_{\infty,1} < \lambda_b\right] \\
&\geq \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{Z}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{W}_j^{(k)}\right| < \lambda_b\right] \\
&\geq \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{Z}_j^{(k)}\right| < t_0\right] \\
&\qquad\qquad \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{W}_j^{(k)}\right| < \lambda_b - t_0\right] \\
&\geq \left(1-2\exp\left(-\frac{t_0^2\left(\sqrt{n}-\sqrt{s}\right)^2}{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)s\lambda_b^2} + r\log(2) + \log\left(p-(2-\alpha)s\right)\right)\right) \\
&\qquad\qquad \left(1-2\exp\left(-\frac{(\lambda_b-t_0)^2 n}{4\sigma^2 r} + r\log(2) + \log\left(p-(2-\alpha)s\right)\right)\right).
\end{aligned}$$

This probability goes to 1 for

$$t_0 = \frac{\sqrt{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)ns}\lambda_b}{\sqrt{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)ns}\lambda_b + 2\sigma(\sqrt{n}-\sqrt{s})}\lambda_b$$

(the solution to $\frac{(\lambda_b - t_0)^2 n}{4\sigma^2 r} = \frac{t_0^2(\sqrt{n}-\sqrt{s})^2}{\frac{1}{\kappa^2}(f_1(\kappa)+f_2(\kappa))s\lambda_b^2}$), if

$$\lambda_b > \frac{\sqrt{4\sigma^2\left(1-\sqrt{\frac{s}{n}}\right)^2 r\Big(r\log(2)+\log\big(p-(2-\alpha)s\big)\Big)}}{\sqrt{n}-\left(\sqrt{s}+\sqrt{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)sr\Big(r\log(2)+\log\big(p-(2-\alpha)s\big)\Big)}\right)}$$

provided that (substituting $r=2$),

$$\begin{aligned} n > {} & \frac{2}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)s\log\Big(p-(2-\alpha)s\Big) \\ & + \Bigg(1+\frac{2}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)2\log(2) \\ & \quad + 2\sqrt{\frac{2}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)\Big(2\log(2)+\log\Big(p-(2-\alpha)s\Big)\Big)}\Bigg)s. \end{aligned}$$

For large enough p with $\frac{s}{p} = \mathbf{o}(1)$, we require

$$n > \frac{2}{\kappa^2} f(\kappa) s \log\Big(p-(2-\alpha)s\Big).$$

Combining this result with (2.7), the lemma follows.

□

2.7 Deterministic Necessary Optimality Conditions

In this appendix, we investigate deterministic necessary conditions for the optimality of the solutions $(\hat{B}, \hat{S})$ of the problem (2.1).

2.7.1 Sub-differential of ℓ_1/ℓ_∞ and ℓ_1/ℓ_1 Norms

In this section we state the sub-differential characterization of the norms we used in out convex program. The results can be directly derived from the definition of sub-differential of a function.

Lemma 9 (Sub-differential of ℓ_1/ℓ_∞-Norm)**.** *The matrix $\widetilde{Z} \in \mathbb{R}^{p\times r}$ belongs to the sub-differential of ℓ_1/ℓ_∞-norm of matrix $\widetilde{B}$, denoted as $\widetilde{Z} \in \partial \left\|\widetilde{B}\right\|_{1,\infty}$ iff*

(i) for all $j \in RowSupp(\widetilde{B})$, we have $\tilde{z}_j^{(k)} = \begin{cases} t_j^{(k)}\, sign\left(\tilde{b}_j^{(k)}\right) & k \in M_j(\widetilde{B}) \\ 0 & ow. \end{cases}$, where, $t_j^{(k)} \geq 0$ and $\sum_{k=1}^r t_j^{(k)} = 1$.

(ii) for all $j \notin RowSupp(\widetilde{B})$, we have $\sum_{k=1}^r \left|\tilde{z}_j^{(k)}\right| \leq 1$.

Lemma 10 (Sub-differential of ℓ_1/ℓ_1-Norm)**.** *The matrix $\widetilde{Z} \in \mathbb{R}^{p\times r}$ belongs to the sub-differential of ℓ_1/ℓ_1-norm of matrix $\widetilde{S}$, denoted as $\widetilde{Z} \in \partial \left\|\widetilde{S}\right\|_{1,1}$ iff*

(i) for all $(j,k) \in Supp(\widetilde{S})$, we have $\tilde{z}_j^{(k)} = sign\left(\tilde{s}_j^{(k)}\right)$.

(ii) for all $(j,k) \notin Supp(\widetilde{S})$, we have $\left|\tilde{z}_j^{(k)}\right| \leq 1$.

2.7.2 Necessary Conditions

The first lemma shows a necessary condition for any solution of the problem (2.1).

Lemma 11. *If $(\hat{S}, \hat{B})$ is a solution (uniqueness is NOT required) of (2.1) then the following properties hold*

(P1) $sign(\hat{s}_j^{(k)}) = sign(\hat{b}_j^{(k)})$ for all $(j,k) \in Supp(\hat{S})$ with $j \in RowSupp(\hat{B})$.

(P2) if $\frac{\lambda_b}{\lambda_s}$ is not an integer, $\frac{1}{D(\hat{S})} > \frac{\lambda_s}{\lambda_b} > \frac{1}{M(\hat{B})}$.

(P3) $\left|\hat{b}_j^{(k)}\right| = \left\|\hat{b}_j\right\|_\infty$ for all $(j,k) \in Supp(\hat{S})$.

(P4) if $\frac{\lambda_b}{\lambda_s}$ is not an integer, $\forall j\, \exists k$ such that $(j,k) \notin Supp(\hat{S})$ and $\left|\hat{b}_j^{(k)}\right| = \left\|\hat{b}_j\right\|_\infty$.

Proof. We provide the proof of each property separately.

(P1) Suppose there exists $(j_0, k_0) \in \text{Supp}(\hat{S})$, such that $\text{sign}(\hat{s}_j^{(k)}) = -\text{sign}(\hat{b}_j^{(k)})$. Let $\check{B}, \check{S} \in \mathbb{R}^{p\times r}$ be matrices equal to $\hat{B}, \hat{S}$ in all entries except at (j_0, k_0). Consider the following two cases

1. $\left|\hat{s}_{j_0}^{(k_0)} + \hat{b}_{j_0}^{(k_0)}\right| \leq \left\|\hat{b}_{j_0}\right\|_\infty$: Let $\check{b}_{j_0}^{(k_0)} = \hat{b}_{j_0}^{(k_0)} + \hat{s}_{j_0}^{(k_0)}$ and $\check{s}_{j_0}^{(k_0)} = 0$. Notice that $(j_0, k_0) \notin \text{Supp}(\check{S})$.
2. $\left|\hat{s}_{j_0}^{(k_0)} + \hat{b}_{j_0}^{(k_0)}\right| > \left\|\hat{b}_{j_0}\right\|_\infty$: Let $\check{b}_{j_0}^{(k_0)} = -\text{sign}\left(\hat{b}_{j_0}^{(k_0)}\right)\left\|\hat{b}_{j_0}\right\|_\infty$ and $\check{s}_{j_0}^{(k_0)} = \hat{s}_{j_0}^{(k_0)} + \hat{b}_{j_0}^{(k_0)} - \check{b}_{j_0}^{(k_0)}$. Notice that $\text{sign}\left(\check{b}_{j_0}^{(k_0)}\right) = \text{sign}\left(\check{s}_{j_0}^{(k_0)}\right)$.

Since $\check{B} + \check{S} = \hat{B} + \hat{S}$ and $\|\check{b}_{j_0}\|_\infty \leq \|\hat{b}_{j_0}\|_\infty$ and $\|\check{s}_{j_0}\|_1 < \|\hat{s}_{j_0}\|_1$, it is a contradiction to the optimality of $(\hat{B}, \hat{S})$.

(P2) We prove the result in two steps by establishing 1. $M(\hat{B}) > \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor$ and 2. $D(\hat{S}) < \left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil$.

1. In contrary, suppose there exists a row $j_0 \in \text{RowSupp}(\hat{B})$ such that $\left|M_{j_0}(\hat{B})\right| \leq \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor$. Let k^* be the index of the element whose magnitude is ranked $\left(\left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor + 1\right)$ among the element of the vector $\hat{b}_{j_0} + \hat{s}_{j_0}$. Let $\check{B}, \check{S} \in \mathbb{R}^{p\times r}$ be matrices equal to $\hat{B}, \hat{S}$ in all entries except on the row j_0 and

$$\hat{b}_{j_0}^{(k)} = \begin{cases} \left|\hat{b}_{j_0}^{(k^*)} + \hat{s}_{j_0}^{(k^*)}\right| \text{sign}\left(\hat{b}_{j_0}^{(k)}\right) \\ \qquad\qquad \left|\hat{b}_{j_0}^{(k)} + \hat{s}_{j_0}^{(k)}\right| \geq \left|\hat{b}_{j_0}^{(k^*)} + \hat{s}_{j_0}^{(k^*)}\right| \\ \hat{b}_{j_0}^{(k)} + \hat{s}_{j_0}^{(k)} \qquad\qquad \text{ow,} \end{cases}$$

and $\check{s}_{j_0} = \hat{s}_{j_0} + \hat{b}_{j_0} - \check{b}_{j_0}$. Notice that $M(\check{B}) > \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor$ and $\text{sign}\left(\check{s}_{j_0}^{(k)}\right) = \text{sign}\left(\check{b}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in \text{Supp}\left(\check{s}_{j_0}\right)$ since $\text{sign}\left(\hat{s}_{j_0}^{(k)}\right) = \text{sign}\left(\hat{b}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in \text{Supp}\left(\hat{S}_{j_0}\right)$ by (P1). Further, since $\check{S} + \check{B} = \hat{S} + \hat{B}$ and $\|\check{b}_{j_0}\|_\infty = \left|\hat{b}_{j_0}^{(k^*)}\right| + \left|\hat{s}_{j_0}^{(k^*)}\right|$ and $\|\check{s}_{j_0}\|_1 \leq \|\hat{s}_{j_0}\|_1 + \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor \left(\left\|\hat{b}_{j_0}\right\|_\infty - \left|\check{b}_{j_0}^{(k^*)}\right| - \left|\check{s}_{j_0}^{(k^*)}\right|\right)$,

this is a contradiction to the optimality of $(\hat{B}, \hat{S})$ due to the fact that $\lambda_s \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor < \lambda_b$.

2. In contrary, suppose there exists a row $j_0 \in \text{RowSupp}(\hat{S})$ such that $\|\hat{s}_{j_0}\|_0 \geq \left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil$. Let k^* be the index of the element whose magnitude is ranked $\left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil$ among the elements of the vector $\hat{b}_{j_0} + \hat{s}_{j_0}$. Let $\check{B}, \check{S} \in \mathbb{R}^{p \times r}$ be matrices respectively equal to $\hat{B}$ and $\hat{S}$ in all entries except on the row j_0 and

$$\hat{b}_{j_0}^{(k)} = \begin{cases} \left|\hat{b}_{j_0}^{(k^*)} + \hat{s}_{j_0}^{(k^*)}\right| \text{sign}\left(\hat{b}_{j_0}^{(k)}\right) \\ \qquad\qquad\qquad \left|\hat{b}_{j_0}^{(k)} + \hat{s}_{j_0}^{(k)}\right| \geq \left|\hat{b}_{j_0}^{(k^*)} + \hat{s}_{j_0}^{(k^*)}\right| \\ \hat{b}_{j_0}^{(k)} + \hat{s}_{j_0}^{(k)} \qquad\qquad\qquad \text{ow,} \end{cases}$$

and $\check{s}_{j_0} = \hat{s}_{j_0} + \hat{b}_{j_0} - \check{b}_{j_0}$. Notice that $D(\check{S}) < \left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil$ and $\text{sign}\left(\check{s}_{j_0}^{(k)}\right) = \text{sign}\left(\check{b}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in \text{Supp}\left(\check{s}_{j_0}\right)$ since $\text{sign}\left(\hat{s}_{j_0}^{(k)}\right) = \text{sign}\left(\hat{b}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in \text{Supp}\left(\hat{s}_{j_0}\right)$. Since $\check{S} + \check{B} = \hat{S} + \hat{B}$ and $\|\check{b}_{j_0}\|_\infty = \left|\hat{b}_{j_0}^{(k^*)}\right| + \left|\hat{s}_{j_0}^{(k^*)}\right|$ and $\|\check{s}_{j_0}\|_1 \leq \|\hat{s}_{j_0}\|_1 + \left(\left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil - 1\right)\left(\left\|\hat{b}_{j_0}\right\|_\infty - \left|\check{b}_{j_0}^{(k^*)}\right| - \left|\check{s}_{j_0}^{(k^*)}\right|\right)$, this is a contradiction to the optimality of $(\hat{B}, \hat{S})$, due to the fact that $\lambda_s \left(\left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil - 1\right) < \lambda_s \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor < \lambda_b$.

(P3) If $j \notin \text{RowSupp}(\hat{B})$ then the result is trivial. Suppose there exists $(j_0, k_0) \in \text{Supp}(\hat{S})$ with $j_0 \in \text{RowSupp}(\hat{S})$ such that $\left|b_{j_0}^{(k_0)}\right| < \|\hat{b}_{j_0}\|_\infty$. Let $\check{B}, \check{S} \in \mathbb{R}^{p \times r}$ be matrices equal to $\hat{B}, \hat{S}$ in all entries except for the entry corresponding to the index (j_0, k_0). Let $\check{b}_{j_0}^{(k_0)} = \left\|\hat{b}_{j_0}\right\|_\infty \text{sign}\left(\hat{b}_{j_0}^{(k_0)}\right)$ if $\left|\hat{b}_{j_0}^{(k_0)} + \hat{s}_{j_0}^{(k_0)}\right| \geq \|b_{j_0}\|_\infty$ and $\check{b}_{j_0}^{(k_0)} = \hat{b}_{j_0}^{(k_0)} + \hat{s}_{j_0}^{(k_0)}$ otherwise. Let $\check{s}_{j_0}^{(k_0)} = \hat{s}_{j_0}^{(k_0)} + \hat{b}_{j_0}^{(k_0)} - \check{b}_{j_0}^{(k_0)}$. Since $\check{B} + \check{S} = \hat{B} + \hat{S}$ and $\left\|\check{b}_{j_0}\right\|_\infty = \left\|\hat{b}_{j_0}\right\|_\infty$ and $\|\check{s}_{j_0}\|_1 < \|\hat{s}_{j_0}\|_1$, it is a contradiction to the optimality of $(\hat{B}, \hat{S})$.

(P4) If $j \notin \text{RowSupp}(\hat{B})$ or $j \notin \text{RowSupp}(\hat{S})$ the result is trivial. Suppose there exists a row $j_0 \in \text{RowSupp}(\hat{B}) \cap \text{RowSupp}(\hat{S})$ such that the result

does not hold for that. Let $k^* = \arg\max_{\{k:(j,k)\notin \mathrm{Supp}(\hat{S})\}} \left|\hat{b}_j^{(k)}\right|$. Let $\check{B}, \check{S} \in \mathbb{R}^{p\times r}$ be matrices equal to $\hat{B}, \hat{S}$ in all entries except for the row j_0 and

$$\hat{b}_{j_0}^{(k)} = \begin{cases} \left|\hat{b}_{j_0}^{(k^*)}\right| \operatorname{sign}\left(\hat{b}_{j_0}^{(k)}\right) & (j_0,k) \in \mathrm{Supp}(\hat{S}) \\ \hat{b}_{j_0}^{(k)} & \text{ow,} \end{cases}$$

and $\check{s}_{j_0} = \hat{s}_{j_0} + \hat{b}_{j_0} - \check{b}_{j_0}$. Since $\check{B} + \check{S} = \hat{S} + \hat{B}$ and $\left\|\check{b}_{j_0}\right\|_\infty = \left|\hat{b}_{j_0}^{(k^*)}\right|$ and by (P2) and (P3), $\|\check{s}_{j_0}\|_1 \le \|\hat{s}_{j_0}\|_1 + \left(\left\lceil \frac{\lambda_b}{\lambda_s}\right\rceil - 1\right)\left(\left\|\hat{b}_{j_0}\right\|_\infty - \left|\hat{b}_{j_0}^{(k^*)}\right|\right)$, this is a contradiction to the optimality of $(\hat{B}, \hat{S})$, due to the fact that $\lambda_s\left(\left\lceil \frac{\lambda_b}{\lambda_s}\right\rceil - 1\right) < \lambda_s \left\lfloor \frac{\lambda_b}{\lambda_s}\right\rfloor < \lambda_b$.

This concludes the proof of the lemma.

□

The next lemma shows why the assumption that the ratio of penalty regularizer parameters is crucial for our analysis. This is not a deterministic result, but since it is related to optimality conditions, we included this lemma in this appendix.

Lemma 12. *If $(\hat{S}, \hat{B})$ with $\hat{B} \neq \mathbf{0}$ is a solution to (2.1) and $d = \frac{\lambda_b}{\lambda_s}$ is an integer then $(\hat{S}, \hat{B})$ is not the unique solution.*

Proof. In contrary, assume that $(\hat{S}, \hat{B})$ is the unique solution. Take a non-zero row $\hat{b}_{j_0}$ with $j_0 \in \mathrm{RowSupp}(\hat{B})$. If $\left|M_{j_0}(\hat{B})\right| < d$, then let $\check{B}, \check{S} \in \mathbb{R}^{p\times r}$ be two matrices equal to $\hat{B}, \hat{S}$ except on the row j_0 and let $\check{b}_{j_0} = \mathbf{0}$ and $\check{s}_{j_0} = \hat{b}_{j_0} + \hat{s}_{j_0}$. Then, $(\check{B}, \check{S})$ are *strictly* better solutions than $(\hat{B}, \hat{S})$. This contradicts the optimality of $(\hat{B}, \hat{S})$. Hence, $\left|M_{j_0}(\hat{B})\right| \ge d$. with similar argument we can conclude that $\left\|\hat{S}_{j_0}\right\|_0 \le d$.

If $\left\|\hat{S}_{j_0}\right\|_0 = d$, then let $0 < \delta \le \min_{(j_0,k)\in \mathrm{Supp}(\hat{S})} \left|\hat{s}_{j_0}^{(k)}\right|$ and $\check{B}(\delta), \check{S}(\delta) \in \mathbb{R}^{p\times r}$ be two matrices equal to $\hat{B}, \hat{S}$ except for the entries indexed $(j_0, k) \in$

$\text{Supp}(\hat{S})$ and let $\check{b}_{j_0}^{(k)} = \hat{b}_{j_0}^{(k)} + \delta \text{sign}\left(\hat{b}_{j_0}^{(k)}\right)$ and $\check{s}_{j_0}^{(k)} = \hat{s}_{j_0}^{(k)} - \delta \text{sign}\left(\hat{s}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in \text{Supp}(\hat{S})$. Then, $(\check{B}(\delta), \check{S}(\delta))$ is another solution to (2.1). This contradicts the uniqueness of $(\hat{B}, \hat{S})$.

If $\left\|\hat{S}_{j_0}\right\|_0 < d$, then using Lemma 11 and Equation 2.5, we have

$$
\begin{aligned}
&\mathbb{P}\left[\left|M_{j_0}(\hat{B})\right| \geq d+1\right] \\
&= \sum_{i=1}^{r-d} \mathbb{P}\left[\left|M_{j_0}(\hat{B})\right| = d+i\right] \\
&= \sum_{i=1}^{r-d} \mathbb{P}\Bigg[\exists k_1, \ldots, k_{i+1} \in M_{j_0}(\hat{B}) \quad \forall l = 1, \ldots, i+1 : \\
&\qquad\qquad \|\hat{b}_{j_0}^{(k_l)} + \underbrace{\hat{s}_{j_0}^{(k_l)}}_{0}| = \left\|\hat{b}_{j_0}\right\|_\infty\Bigg] \\
&= \sum_{i=1}^{r-d} \mathbb{P}\Bigg[\exists k_1, \ldots, k_{i+1} \in M_{j_0}(\hat{B}) \quad \forall l = 1, \ldots, i+1 : \\
&\qquad\qquad \left|\Delta_{j_0}^{(k_l)}\right| = \left|b_j^{*(k_l)} + s_j^{*(k_l)}\right| + \left\|\hat{b}_j\right\|_\infty\Bigg] \\
&= \sum_{i=1}^{r-d} \mathbb{P}\Bigg[\exists k_1, \ldots, k_{i+1} \in M_{j_0}(\hat{B}) \quad \forall l, m = 1, \ldots, i+1 : \\
&\qquad\qquad \left|\Delta_{j_0}^{(k_l)}\right| = C_{k_l, k_m} + \left|\Delta_{j_0}^{(k_m)}\right|\Bigg] = 0.
\end{aligned}
$$

In above equation C_{k_l,k_m} are some constants. The last conclusion follows from the fact that $\Delta_{j_0}^{(k_l)}$'s are continuous Gaussian variables and the cardinality of this event is less than the cardinality of the space they lie in. Hence, $\left|M_{j_0}(\hat{B})\right| = d$.

Let $0 < \delta < \|b_{j_0}\|_\infty$ and $\check{B}(\delta), \check{S}(\delta) \in \mathbb{R}^{p \times r}$ be two matrices equal to $\hat{B}, \hat{S}$ except for the entries indexed (j_0, k) for $k \in M_{j_0}(\hat{B})$ and let $\check{b}_{j_0}^{(k)} = \hat{b}_{j_0}^{(k)} - \delta$ and $\check{s}_{j_0}^{(k)} = \hat{s}_{j_0}^{(k)} + \delta$ for all $k \in M_{j_0}(\hat{B})$. Then, $(\check{B}(\delta), \check{S}(\delta))$ is another solution to (2.1). This contradicts the uniqueness of $(\hat{B}, \hat{S})$.

□

Next lemma characterizes the optimal solution by introducing a dual

variable $\hat{Z}$.

Lemma 13 (Convex Optimality)**.** *If* $(\hat{B}, \hat{S})$ *is a solution of* (2.1) *then there exists a matrix* $\hat{Z} \in \mathbb{R}^{p \times r}$*, called* dual variable*, such that* $\hat{Z} \in \lambda_s \partial \|\hat{S}\|_{1,1}$ *and* $\hat{Z} \in \lambda_b \partial \|\hat{B}\|_{1,\infty}$ *and for all* $k = 1, \ldots, r$*,*

$$\frac{1}{n} \left\langle X^{(k)}, X^{(k)} \right\rangle \left(\hat{s}^{(k)} + \hat{b}^{(k)} \right) - \frac{1}{n} (X^{(k)})^T y^{(k)} + \hat{z}^{(k)} = 0. \tag{2.8}$$

Proof. The proof follows from the standard first order optimality argument. □

2.8 Coordinate Descent Algorithm

We use the coordinate descendent algorithm described as follows. The algorithm takes the tuple $(X, Y, \lambda_s, \lambda_b, \epsilon, B, S)$ as input, and outputs $(\hat{B}, \hat{S})$. Note that X and Y are given to this algorithm, while B and S are our initial guess or the warm start of the regression matrices. ϵ is the precision parameter which determines the stopping criterion.

We update elements of the sparse matrix S using the subroutine $UpdateS$, and update elements in the block sparse matrix B using the subroutine $UpdateB$, respectively, until the regression matrices converge. The pseudocode is in Algorithm 1 to Algorithm 3.

2.8.1 Correctness of Algorithms

In this algorithm, B is the block sparse matrix and S is the sparse matrix. We alternatively update B and S until they converge. When updating S, we cycle through each element of S while holding all the other elements of S and B unchanged; When updating B, we update each block B_j (the coefficient vector of the j^{th} feature for r tasks) as a whole, while keeping S and other coefficient vector of B fixed.

Algorithm 2 Our Model Solver

Require: X, Y, λ_b, λ_s, B, S and ε
Ensure: $\hat{S}$ and $\hat{B}$

Initialization:
for $j = 1 : p$ **do**
  for $k = 1 : r$ **do**
    $c_j^{(k)} \leftarrow \left\langle X_j^{(k)}, y^{(k)} \right\rangle$
    for $i = 1 : p$ **do**
      $d_{i,j}^{(k)} \leftarrow \left\langle X_i^{(k)}, X_j^{(k)} \right\rangle$
    end for
  end for
end for

Updating:
loop
  $S \leftarrow UpdateS(c; d; \lambda_s; B; S)$
  $B \leftarrow UpdateB(c; d; \lambda_b; B; S)$
  if Relative Update $< \epsilon$ **then**
    BREAK
  end if
end loop
RETURN $\hat{B} = B,\ \hat{S} = S$

Algorithm 3 UpdateB

Require: c, d, λ_b, B and S

Ensure: B

Update B using the cyclic coordinate descent algorithm for ℓ_1/ℓ_∞ while keeping S unchanged.

for $j = 1 : p$ **do**
  for $k = 1 : r$ **do**
    $\alpha_j^{(k)} \leftarrow \left(c_j^{(k)} - \sum_{i \neq j}(b_i^{(k)} + s_i^{(k)})d_{i,j}^{(k)}\right) / d_{j,j}^{(k)} - s_i^{(k)}$
    if $\sum_{k=1}^{r} |\alpha_j^{(k)}| \leq \lambda_b$ **then**
      $b_j \leftarrow 0$
    else
      Sort α to be $|\alpha_j^{(k_1)}| \geq |\alpha_j^{(k_2)}| \geq \cdots \geq |\alpha_j^{(k_r)}|$
      $m^* = \arg\max_{1 \leq m \leq r}(\sum_{k=1}^{r} |\alpha_j^{(k_m)}| - \lambda_b)/m$
      for $i = 1 : r$ **do**
        if $i > m^*$ **then**
          $b_j^{(k_i)} \leftarrow \alpha_j^{(k_i)}$
        else
          $b_j^{(k_i)} \leftarrow \frac{\text{sign}(\alpha_j^{(k_i)})}{m^*} \left(\sum_{l=1}^{m^*} |\alpha_j^{(k_l)}| - \lambda_b\right)$
        end if
      end for
    end if
  end for
end for
RETURN B

Algorithm 4 Update-S

Require: c, d, λ_s, B and S
Ensure: S
  Update S using the cyclic coordinate descent algorithm for LASSO while keeping B unchanged.
  for $j = 1 : p$ **do**
    for $k = 1 : r$ **do**
      $\alpha_j^{(k)} \leftarrow \left(c_j^{(k)} - \sum_{i \neq j}(b_i^{(k)} + s_i^{(k)})d_{i,j}^{(k)}\right) / d_{j,j}^{(k)} - b_i^{(k)}$
      if $|\alpha_j^{(k)}| \leq \lambda_s$ **then**
        $s_j^k \leftarrow 0$
      else
        $s_j^k \leftarrow \alpha_j^{(k)} - \lambda_s \text{sign}(\alpha_j^{(k)})$
      end if
    end for
  end for
  RETURN S

For updating B, the subproblem is updating B_j

$$\hat{b}_j = \arg\min_{b_j} \quad \frac{1}{2}\sum_{k=1}^{r} \left\| r_j^{(k)} - b_j^{(k)} X_j^{(k)} \right\|_2^2 + \lambda_b \|b_j\|_\infty. \tag{2.9}$$

If we take the partial residual vector $r_j^{(k)} = y^{(k)} - \sum_{l \neq j}(b_l^{(k)} X_l^{(k)}) - \sum_l (s_l^{(k)} X_l^{(k)})$, the correctness of this algorithm will directly follow from the correctness of coordinate descent algorithm of ℓ_1/ℓ_{inf} in [88]. With the same argument, the correctness of the Algorithm 3 can be proven.

Chapter 3

Clustering Partially Observed Graphs

This chapter considers the problem of clustering a partially observed unweighted graph – i.e. one where for some node pairs we know there is an edge between them, for some others we know there is no edge, and for the remaining we do not know whether or not there is an edge. We want to organize the nodes into disjoint clusters so that there is relatively dense (observed) connectivity within clusters, and sparse across clusters.

We take a novel yet natural approach to this problem, by focusing on finding the clustering that minimizes the number of "disagreements" - i.e. the sum of the number of (observed) missing edges within clusters, and (observed) present edges across clusters. Our algorithm uses convex optimization; its basis is a reduction of disagreement minimization to the problem of recovering an (unknown) low-rank matrix and an (unknown) sparse matrix from their partially observed sum. We show that our algorithm succeeds under certain natural assumptions on the optimal clustering and its disagreements. While our algorithm is based on matrix spliting technique, because of special property of our problem, our results significantly strengthen existing matrix splitting results and directly enhance solutions to the problem of Correlation Clustering [10] with partial observations.

3.1 Introduction

This chapter is about the following task: given partial observation of an undirected unweighted graph, partition the nodes into disjoint clusters so that there are dense connections within clusters, and sparse connections across clusters. By partial observation, we mean that for some node pairs we know if there is an edge or not, and for other node pairs we do not know – these

pairs are *unobserved*. This problem arises in several fields across science and engineering. For example, in sponsored search, each cluster is a submarket that represents a specific group of advertisers that do most of their spending on a group of query phrases – see e.g. [148] for such a project at Yahoo. In VLSI and design automation, it is useful in minimizing signaling between components, layout etc. – see e.g. [75] and references thereof. In social networks, clusters represent groups of mutual friends; finding clusters enables better recommendations, link prediction, etc [98]. In the analysis of document databases, clustering the citation graph is often an essential and informative first step [47]. In this chapter, we will focus not on specific application domains, but rather on the basic graph clustering problem itself.

As with any clustering problem, this needs a precise mathematical definition of the clustering criterion with potentially a guaranteed performance. We are not aware of any existing work with provable performance guarantees for partially observed graphs. Even most existing approaches to clustering fully observed graphs, which we review in section 3.1.1 below, either require an additional input (e.g. the number of clusters k required for spectral or k-means clustering approaches), or do not guarantee the performance of the clustering. Indeed, the specialization of our results to the fully observed case extends the known guarantees there.

Our Formulation: We focus on a natural formulation, one that *does not require any other extraneous input* besides the graph itself. It is based on minimizing *disagreements*, which we now define. Consider any candidate clustering; this will have *(a)* observed node pairs that are in different clusters, but have an edge between them, and *(b)* observed node pairs that are in the same cluster, but do not have an edge between them. The total number of node pairs of types *(a)* and *(b)* is the number of disagreements between the clustering and the given graph. We focus on the problem of finding the *optimal clustering* – one that minimizes the number of disagreements. Note that we do *not* pre-specify the number of clusters. For the special case of fully observed graphs, this formulation is exactly the same as the problem of "Correlation Clustering", first proposed by [10]. They showed that exact minimization of the above objective is NP-complete in the worst case – we survey and compare this and other related work in section 3.1.1. As we will see, our approach and results are very different.

Our Approach: We aim to achieve the combinatorial disagreement minimization objective using matrix splitting via convex optimization. In particular, as we show in section 2.3 below, one can represent the adjacency matrix of the given graph as the sum of an unknown low-rank matrix (corresponding to"ideal" clusters) and a sparse matrix (corresponding to disagreements from this "ideal" in the given graph). Our algorithm either returns a clustering, which is guaranteed to be disagreement minimizing, or returns a "failure" – it never returns a sub-optimal clustering. Our analysis provides both deterministic and probabilistic guarantees for when our algorithm succeeds. Our analysis uses the special structure of our problem to provide much stronger guarantees than are current results on general matrix splitting [21, 31, 63].

3.1.1 Related Work

Our problem can be interpreted in the general clustering context as one in which the presence of an edge between two points indicates a "similarity", and the lack of an edge means no similarity. The general field of clustering is of course vast, and a detailed survey of all methods therein is beyond our scope here. We focus instead on the two sets of papers most relevant to the problem here, namely the work on Correlation Clustering, and the other approaches to the specific problem of graph clustering.

Correlation Clustering: First formulated in [10], correlation clustering looks at the following problem: given a complete graph where every edge is labelled "+" or "-", cluster the nodes to minimize the total of the number of "-" edges within clusters and "+" edges across clusters. As mentioned, for a completely observed graph, our problem is mathematically precisely the same as correlation clustering; in particular a "+" in correlation clustering corresponds to an edge in graph clustering, and a "-" to the lack of an edge. Disagreements are defined in the same way. Thus, this chapter can equivalently be considered an algorithm, and guarantees, for *correlation clustering under partial observations*. [10] show that exact minimization is NP-complete, and also provide *(a)* constant-factor approximation algorithm for the problem of minimizing the number of disagreements, and *(b)* a PTAS for maximizing agreements. Their

algorithms are combinatorial in nature. Subsequently, there has been much work on devising alternative approximation algorithms for both the weighted and unweighted cases, and for both agreement and disagreement objectives [14, 32, 42, 45, 46, 129]. Approximations based on LP relaxation [14] and SDP relaxation [129], followed by rounding, have also been developed. We emphasize that while we do convex relaxation as well, we do not do rounding; rather, our convex program itself yields an optimal clustering. We emphasize that ours is the *first* attempt at correlation clustering with partial observations. The recent result in [105] has the same flavor under an additional (strong) assumption that the sum of the square of cluster sizes are known apriori, but it requires more observations than our results.

The result in [93] also considers a convex optimization formulation, but with extra constraints including positive semi-definiteness, triangle inequality and fixed diagonal entries. Their guarantee in order-wise identical to ours for the fully-observed, probabilistic case, except that we do not leverage the extra information imposed by those constraints. In short, theoretically our tight analysis provides the exact same guarantee with less constraints and practically our method is faster since, to the best of our knowledge, there is no low-complexity algorithm to deal with positive semi-definite constraint as required by [93]. This means that our method can handle very large graphs while the result of [93] is practically restricted to small graphs (~ 100 nodes). In summary, since they add more constraints, in practice there are likely to be instances where their convex program works and ours does not. But that comes at the expense of much higher computational complexity; also these gains do not seem to be theoretically characterizable.

Graph Clustering: The problem of graph clustering is well studied and very rich literature on the subject exists (see e.g. [48, 66] and references thereof). One set of approaches seek to optimize criteria such as k-median, minimum sum or minimum diameter [17]; typically these result in NP-hard problems with few global guarantees. Another option is a top-down hierarchical approach, i.e., recursively bisecting the graph into smaller and smaller clusters. Various algorithms in this category differ in the criterion used to

determine where to split in each iteration. Notable examples of such criteria include small cut [39], maximal flow [52], low conductance [120], eigenvector of the Laplacian (aka spectral clustering) [103], and many others. Due to the iterative nature of these algorithms, global theoretical guarantees are hard to obtain.

As we mentioned before, we are not aware of any work on graph clustering with partial observations and provable guarantees. However, the nuclear norm minimization has been used to solve planted clique problem [2, 5].

3.2 Main Contributions

Our algorithm is based on convex optimization, and either *(a)* outputs a clustering that is guaranteed to be the one that minimizes the number of observed disagreements, or *(b)* declares "failure" – in which case one could potentially try some other approximate methods. In particular, it never produces a suboptimal clustering. We now briefly present the main idea, then describe the algorithm, and finally present our main results – analytical characterizations of when the algorithm succeeds.

Setup: We are given a partially observed graph, whose adjacency matrix is $\mathbf{A}$ – which has $a_{ij} = 1$ if there is an edge between nodes i and j, $a_{ij} = 0$ if there is no edge, and $a_{ij} =?$ if we do not know. (Here we follow the convention that $a_{ii} = 0$ for all i.) Let Ω_{obs} be the set of observed entries, i.e. the set of elements of $\mathbf{A}$ that are known to be 0 or 1. We want to find the *optimal clustering*, i.e. the one that has the minimum number of disagreements in Ω_{obs}.

Idea: Consider first the fully observed case, i.e. every $a_{ij} = 0$ or 1. Suppose also that the graph is already ideally clustered – i.e. there is a partition of the nodes such that there are no edges between partitions, and each partition is a clique. In this case, the matrix $\mathbf{A} + \mathbf{I}$ is now a *low-rank* matrix, with the rank equal to the number of clusters. This can be seen by noticing that if we re-ordered the rows and columns so that partitions appear together, the result would be a *block-diagonal* matrix, with each block being an all-ones sub-matrix – and thus rank one. Of course, this re-ordering does not change

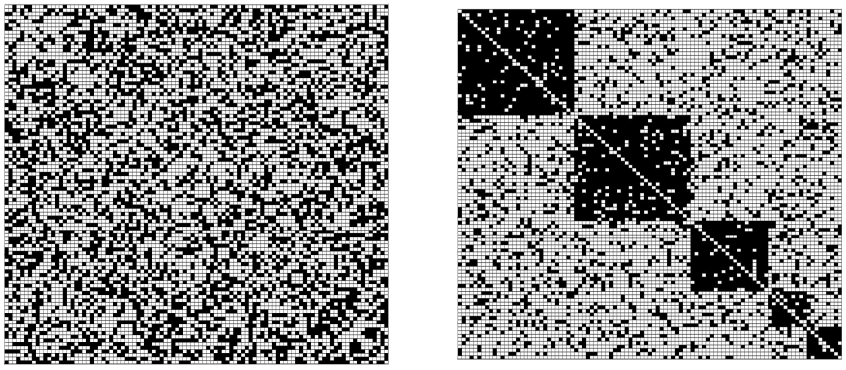


Figure 3.1: The adjacency matrix of a graph before (a) and after (b) proper reordering (i.e. clustering) of the nodes. The figure (b) is indicative of the matrix as a superposition of a sparse matrix and a low-rank one.

the rank of the matrix, and hence $\mathbf{A} + \mathbf{I}$ is (exactly) low-rank.

Consider now any given graph, still fully observed. In light of the above, we are looking for a decomposition of its $\mathbf{I} + \mathbf{A}$ into a low-rank part $\mathbf{K}^*$ (of block-diagonal all-ones, one block for each cluster) and a remaining $\mathbf{B}^*$ (the disagreements) – such that the number of entries in $\mathbf{B}^*$ is as small as possible; i.e. $\mathbf{B}^*$ is sparse. Finally, the problem we look at is recovery of the best $\mathbf{K}^*$ when we do not observe all entries. The idea is depicted in Figure 3.1.

Convex Optimization Formulation: We propose to do the matrix splitting using convex optimization, an approach recently taken in [21, 31, 34]; however, we establish much stronger results for our special problem. Our approach consists of *dropping* any additional structural requirements, and just looking for a decomposition of the given $\mathbf{A}+\mathbf{I}$ as the sum of a sparse matrix $\mathbf{B}$

and a low-rank matrix $\mathbf{K}$. In particular, we use the following convex program

$$\begin{aligned} \min_{\mathbf{B},\mathbf{K}} \quad & \eta \, ||\mathbf{B}||_1 + (1-\eta) \, ||\mathbf{K}||_* \\ \text{s.t.} \quad & \mathcal{P}_{\Omega_{obs}}(\mathbf{B}+\mathbf{K}) = \mathcal{P}_{\Omega_{obs}}(\mathbf{I}+\mathbf{A}) \end{aligned} \tag{3.1}$$

Here, for any matrix M, the term $\mathcal{P}_{\Omega_{obs}}(M)$ keeps all elements of M in Ω_{obs} unchanged, and sets all other elements to 0; the constraints thus state that the sparse and low-rank matrix should in sum be consistent with the observed entries. $||\mathbf{B}||_1 = \sum_{i,j} |b_{ij}|$ is the ℓ_1 norm of the entries of the matrix, which is well-known to be a convex surrogate for the number of non-zero entries $||\mathbf{B}||_0$. The second term is $||\mathbf{K}||_* = \sum_s \sigma_s(K)$ is "nuclear norm": the sum of singular values of $\mathbf{K}$. This has been shown recently to be the convex surrogate[1] for the rank function [115]. Thus our objective function is a convex surrogate for the (natural) combinatorial objective $\eta \, ||\mathbf{B}||_0 + (1-\eta)\text{rank}(\mathbf{K})$. (3.1) is, in fact, a semi-definite program (SDP) [31].

Definitition: *Validity:* The convex program (3.1) is said to produce a *valid* output if the low-rank matrix part $\mathbf{K}$ of the optimum corresponds to a graph of disjoint cliques; i.e. its rows and columns can be re-ordered to yield a block-diagonal matrix with all-one matrices for each block.

Validity of a given $\mathbf{K}$ can easily be checked, either via elementary re-ordering operations, or via a singular value decomposition[2]. Our first simple, but crucial, insight is that whenever the convex program (3.1) yields a valid solution, it is the disagreement minimizer.

Theorem 5. *For any $\eta > 0$, if the optimum of* (3.1) *is valid, then it is the clustering that minimizes the number of observed disagreements.*

Algorithm: Our algorithm takes the adjacency matrix of the network $\mathbf{A}$ and outputs either the optimal clustering or declares failure. Using the

[1] In particular, it is the ℓ_1 norm of the singular value vector, while rank is the ℓ_0 norm of the same.

[2] An SVD of a valid $\mathbf{K}$ will yield singular vectors with disjoint supports. The supports correspond to the clusters.

result of Theorem 5, if the clustering is valid, then we are guaranteed that the result is a disagreement minimizer clustering.

Algorithm 5 Optimal-Cluster(A)

```
for η ∈ (0, 1) do
  Solve (3.1)
  if Solution K is valid then
    Output the clustering w.r.t K and EXIT.
  end if
end for
Declare Failure.
```

We recommend using the fast implementation algorithms developed in [84], which is specially tailored for matrix splitting. Setting the parameter η can be done either via a simple line search from 0 to 1, binary search, or any other option. Whenever it results in a valid $\mathbf{K}$, we have found the optimal clustering.

Analysis: The main analytical contribution of this paper is conditions under which the above algorithm will find the clustering that minimizes the number of disagreements among the observed entries. We provide both deterministic/worst-case guarantees, and average case guarantees for a natural randomness assumption. Let $\mathbf{K}^*$ be the low-rank matrix corresponding to the optimal clustering (as described above). Let $\mathbf{B}^* = \mathcal{P}_{\Omega_{\text{obs}}}(\mathbf{A} + \mathbf{I} - \mathbf{K})$ be the matrix of observed disagreements for this clustering. Note that the support of $\mathbf{B}^*$ is contained in Ω_{obs}. Let $K_{\min}$ be the size of the smallest cluster in $\mathbf{K}^*$.

Deterministic guarantees: We first provide deterministic conditions under which (3.1) will find $\mathbf{K}^*$. For any node i, let $C(i)$ be the cluster in $\mathbf{K}^*$ that node i belongs to. For any cluster $c \neq C(i)$, define $d_{i,c} = |\{j \in c \,|\, a_{ij} = ? \text{ or } a_{ij} = 1\}|$ and for $c = C(i)$, define $d_{i,c} = |\{j \in c \,|\, a_{ij} = ? \text{ or } a_{ij} = 0\}|$. In words, for both cases, $d_{i,c}$ is the total number of disagreements and unobserved entries between i and c. We now define a quantity $D_{\max}$ as follows

$$D_{\max} \;=\; \max_{i,c} \frac{d_{i,c}}{\min\{|c|, C(i)\}}$$

Essentially, $D_{\max}$ is the largest *fraction* of "bad entries" (i.e. disagreements or unobserved) between a node and a cluster. Thus for the same $D_{\max}$, a node is

allowed to have more bad entries to a larger cluster, but constrained to have a smaller to a smaller cluster. It is intuitively clear that a large D_{max} will cause problems, as a node will have so many disagreements (with respect to the corresponding cluster size) that it will be impossible to resolve. We now state our main theorem for the deterministic case.

Theorem 6. *If* $\frac{nD_{\max}}{K_{\min}} < \frac{1}{4}$*, then the optimal clustering* $(\mathbf{K}^*, \mathbf{B}^*)$ *is the unique solution of (3.1) for any*

$$\eta \in \left(\frac{1}{1 + \frac{1}{2}K_{\min}} \, , \, 1 - \frac{K_{\min}}{\left(1 + \frac{3}{4nD_{\max}}\right) K_{\min} - 1} \right).$$

Remarks on Theorem 6: Essentially, Theorem 6 allows for the number of disagreements and unobserved edges at a node to be as large as a third of the number of "good" edges (i.e. edges to its own cluster in the optimal clustering). This means that there is a lot of evidence "against" the optimal clustering, and missing evidence, making it that much harder to find. Theorem 6 allows a node to have many disagreements and unobserved edges overall; it just requires these to be distributed proportional to the cluster sizes.

In many applications, the size of the typical cluster may be much smaller than the size of the graph. Theorem 6 implies that the smallest cluster $K_{\min} > 4\sqrt{n}$ for any non-trivial problem (i.e. one where every cluster has at least one node with at least one disagreement or unobserved edge). Our method can thus handle as many as $\Theta(\sqrt{n})$ clusters; this can be compared to existing approaches to graph clustering, which often partition nodes into two or a constant number of clusters. The guarantees of this theorem are orderwise stronger than what would result from a direct application of the deterministic guarantees in [31, 63]. Indeed, the results in [63] implies correct recovery as long as $D_{\max} \leq c\frac{K_{\min}^2}{n^2}$ for some constant c. (This result subsumes those in [31].) Theorem 6 only requires $D_{\max} < \frac{K_{\min}}{4n}$, which is an order improvement if $K_{\min}$ grows slower than n.

Probabilistic Guarantees: We now provide much stronger guarantees for the case where both the locations of the observations, and the locations of the observed disagreements, are drawn uniformly at random. Specifically, consider a graph that is generated as follows: start with an initial "ideally clustered" graph with no disagreements – i.e. each cluster is completely connected (i.e. a full clique), and different clusters are completely disconnected (i.e. have no edges between them). Then for some $0 < \tau < 1$ and for each of the $\binom{n}{2}$ possible node pairs, flip the entry in this location with probability τ from 0 to 1 or 1 to 0, as the case may be – thus causing them to be disagreements. There are thus, on average, $\tau\binom{n}{2}$ disagreements in the resulting graph. The actual number is close to this with high probability, by standard concentration arguments. Further, this graph is observed at locations chosen uniformly at random. Specifically, for each node pair (i, j) there is a probability p_0 that $(i, j) \in \Omega_{\text{obs}}$, and this choice is made independently of any other node pair, or of the graph. Note that now it may be possible that the optimal clustering is not the original ideal clustering we started with; the following theorem says that we will still find the optimal clustering with high probability.

Theorem 7. *For any constant $c > 0$, there exist constants C_d, C_k, such that, with probability at least $1 - cn^{-10}$, the optimal clustering $(\mathbf{K}^*, \mathbf{B}^*)$ is the unique solution of (3.1) with $\eta = \frac{1}{1+\sqrt{np_0}}$ provided that*

$$\tau \leq C_d \qquad \textit{and} \qquad K_{\min} \geq C_k \sqrt{n(\log n)^6/p_0}.$$

Remarks on Theorem 7: This shows that our algorithm will succeed in the overwhelming majority of instances where as large as a constant fraction of all observations are disagreements. In particular the number of disagreements can be an order of magnitude larger than the number of "good" edges (i.e. those that agree with the clustering). This remains true even if we observe a vanishingly small fraction of the total number of node pairs – p_0 above is allowed to be a function of n. Smaller p_0 however requires $K_{\min}$ to be correspondingly larger. The reason underlying these stronger results is that bounded matrices with random supports are very spectrally diffuse, and thus find it hard to "hide" a clique, which is highly structured. When p_0 is a constant, our theorem and the probabilistic guarantees in [21] can both handle

the same value of corrupted fraction τ. However, our theorem goes beyond [21] as we allow p_0 to be a vanishing function of n.

Remarks on Outliers: Our algorithm has the capability to handle outliers (i.e., isolated nodes outside the true clusters with at most $D_{\max}|c|$ edges to each true cluster c) by classifying all their edges as disagreements - and hence automatically revealing each outlier as a single-node cluster. In the output of our algorithm, the low rank part $\mathbf{K}$ will have all zeroes in columns corresponding to outliers – all their edges will appear in the disagreement matrix $\mathbf{B}$.

3.3 Proofs

3.3.1 Proof of Theorem 5

In this section, we prove Theorem 5; in particular, that if (3.1) produces a valid low-rank matrix, i.e. one that corresponds to a clustering of the nodes, then this is the disagreement minimizing clustering. Consider the following non-convex optimization problem

$$\begin{aligned} \min_{\mathbf{B},\mathbf{K}} \quad & \eta \, ||\mathbf{B}||_1 + (1-\eta) \, ||\mathbf{K}||_* \\ \text{s.t.} \quad & P_{\Omega_{\text{obs}}}(\mathbf{B}+\mathbf{K}) = P_{\Omega_{\text{obs}}}(\mathbf{I}+\mathbf{A}) \\ & \mathbf{K} \text{ is valid} \end{aligned} \tag{3.2}$$

and let $(\mathbf{B}, \mathbf{K})$ be any feasible solution. Since $\mathbf{K}$ represents a valid clustering, it is positive semidefinite and has all ones along its diagonal. Therefore, any valid $\mathbf{K}$ obeys $\|\mathbf{K}\|_* = \text{trace}(\mathbf{K}) = n$. On the other hand, because both $\mathbf{K}-\mathbf{I}$ and $\mathbf{A}$ are adjacency matrices, the entries of $\mathbf{B} = \mathbf{I}+\mathbf{A}-\mathbf{K}$ must be equal to -1, 1 or 0 (i.e. it is a disagreement matrix). Hence $\|\mathbf{B}\|_1 = \|\mathbf{B}\|_0$ when $\mathbf{K}$ is valid. We thus conclude that the above optimization problem is equivalent to minimizing $||\mathbf{B}||_0$ s.t. the constraints in (3.2) hold. This is exactly the minimization of the number of disagreements on the observed edges. Now notice that (3.1) is a relaxed version (3.2). Therefore, if the optimum of (3.1) is valid and feasible to (3.2), then it is also optimal to (3.2).

3.3.2 Proof Outline for Theorem 6 and 7

We now overview the main steps in the proof of Theorem 6 and 7; the following sections provide details. Recall that we would like to show that $\mathbf{K}^*$ and $\mathbf{B}^*$ corresponding to the optimal clustering is the unique optimum of our convex program (3.1). This involves the following steps:

Step 1: Write down sub-gradient based first-order sufficient conditions that need to be satisfied for $\mathbf{K}^*, \mathbf{B}^*$ to be the unique optimum of (3.1). In our case, this involves showing the existence of a matrix $\mathcal{Q}$ – the *dual certificate* – that satisfies certain properties. This step is technically involved – requiring us to delve into the intricacies of sub-gradients since our convex function is not smooth – but otherwise standard. Luckily for us, this has been done by [21, 31].

Step 2: Using the assumptions made on the optimal clustering and its disagreements $(\mathbf{K}^*, \mathbf{B}^*)$, construct a candidate dual certificate $\mathcal{Q}$ that meets the requirements – and thus certifies $\mathbf{K}^*, \mathbf{B}^*$ as being the unique optimum. This is where the "art" of the proof lies: different assumptions on the $\mathbf{K}^*, \mathbf{B}^*$ (e.g. we look at deterministic and random assumptions) and different ways to construct this $\mathcal{Q}$ will result in different performance guarantees.

The crucial second step is where we go beyond the existing literature on matrix splitting [21, 31]. In particular, our sparse and low-rank matrices have a lot of additional structure, and we use some of this in new ways to generate dual certificates. This leads to much more powerful performance guarantees than those that could be obtained via a direct application of existing sparse and low-rank matrix splitting results. Next, we introduce some notations used in the rest of the paper for the proofs of the theorems.

3.3.2.1 Preliminaries

Definitions related to $\mathbf{K}^*$: By symmetry, the SVD of $\mathbf{K}^*$ is of the form $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$. We define the sub-space $\mathcal{T} = \left\{\mathbf{U}\mathbf{X}^T + \mathbf{Y}\mathbf{U}^T : \mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times p}\right\}$ to be the span of all matrices that share either the same column space or the same row space as $\mathbf{K}^*$. For any matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, we can define its *orthogonal projection* to the space $\mathcal{T}$ as $\mathcal{P}_{\mathcal{T}}(\mathbf{M}) = \mathbf{U}\mathbf{U}^T\mathbf{M} + \mathbf{M}\mathbf{U}\mathbf{U}^T - \mathbf{U}\mathbf{U}^T\mathbf{M}\mathbf{U}\mathbf{U}^T$. We also define the projection onto $\mathcal{T}^\perp$, the complement orthogonal space of $\mathcal{T}$, as

$\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{M}) \;=\; \mathbf{M} - P_{\mathcal{T}}(\mathbf{M})$.

Definitions related to $\mathbf{B}^*$: For any matrix $\mathbf{M}$, define $supp(\mathbf{M}) = \{(i,j) : M_{i,j} \neq 0\}$. Let $\Omega = \{\mathbf{B} \in \mathbb{R}^{n\times n} : supp(\mathbf{B}) \subseteq supp(\mathbf{B}^*)\}$ be the space of matrices with support sets that are a subset of the support set of $\mathbf{B}^*$. Let $\mathcal{P}_\Omega(\mathbf{N}) \in \mathbb{R}^{n\times n}$ be the orthogonal projection of the matrix $\mathbf{N}$ onto the space Ω, i.e., $\mathcal{P}_\Omega(\mathbf{N})$ is obtained from $\mathbf{N}$ by setting all entries not in the set $supp(\mathbf{B}^*)$ to zero. Let $\Omega^\perp$ be the orthogonal space to Ω – it is the space of all matrices whose entries in the set $supp(\mathbf{B}^*)$ are zero. The projection $\mathcal{P}_{\Omega^\perp}$ is defined accordingly. Finally, let $sign(\mathbf{B}^*)$ be the matrix whose entries are +1 for every positive entry in $\mathbf{B}^*$, -1 for every negative entry, and 0 for all the zero entries.

Definitions related to partial observations: Let Ω_{obs} be the space of matrices with support sets that are a subset of the set of observed entries, and $\Gamma = \Omega^\perp \cap \Omega_{\text{obs}}$ is the set of matrices with support within the set of observed entries but outside the set of disgreements. Accordingly, define $\mathcal{P}_{\Omega_{\text{obs}}}$, $\mathcal{P}_{\Omega^\perp_{\text{obs}}}$, $\mathcal{P}_\Gamma$ and $\mathcal{P}_{\Gamma^\perp}$ similar to that of $\mathcal{P}_\Omega$ and $\mathcal{P}_{\Omega^\perp}$.

Norms: $\|\mathbf{M}\|$ and $\|\mathbf{M}\|_F$ represent the spectral and Frobenius norm of the matrix $\mathbf{M}$ respectively and $\|\mathbf{M}\|_\infty = \max_{i,j}|M_{i,j}|$.

3.4 Worst Case Analysis

In this section, we prove Theorem 6. We first state the determinisic first-order conditions required for $\mathbf{B}^*$ and $\mathbf{K}^*$ to be the unique optimum of our convex program (3.1).

Lemma 14 (Deterministic Sufficient Optimality)**.** *[31]* $\mathbf{B}^*$ *and* $\mathbf{K}^*$ *are unique solutions to* (3.1) *provided that* $\mathcal{T} \cap \Gamma^\perp = \{\mathbf{0}\}$ *and there exists a matrix* $\mathcal{Q}$ *such that*

(a). $\mathcal{P}_{\Omega^\perp_{obs}}(\mathcal{Q}) = 0$;
(b). $\mathcal{P}_{\mathcal{T}}(\mathcal{Q}) = (1-\eta)\mathbf{U}\mathbf{U}^T$;
(c). $\mathcal{P}_\Omega(\mathcal{Q}) = \eta sign(\mathbf{B}^*)$;
(d). $\|\mathcal{P}_{\mathcal{T}^\perp}(\mathcal{Q})\| < 1-\eta$;
(e). $\|\mathcal{P}_{\Omega^\perp}(\mathcal{Q})\|_\infty < \eta$.

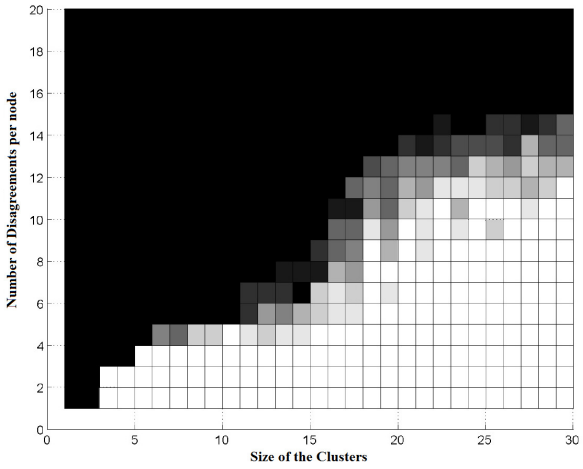


Figure 3.2: Simulation results for fully observed 1000-node graph with all clusters of the same size. For different cluster sizes $K_{\min}$ and different number of disagreements per node b, we plot the probability of success.

The first condition, $\mathcal{T} \cap \Gamma^{\perp} = \{\mathbf{0}\}$, is satisfied under the assumption of the theorem; the proof follows from showing $\|\mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}))\|_{\infty} < \|\mathbf{N}\|_{\infty}$. Next, we need to construct a suitable dual certificate $\mathcal{Q}$ that satisfies condition (a)-(e). We use the alternating projection method (see [22]) to construct $\mathcal{Q}$. The novelty of our analysis is that by taking advantage of the rich structures of the matrices $\mathbf{B}^*$ and $\mathbf{K}^*$, such as symmetricity, block-diagonal, etc, we improve the existing guarantees [21, 31] to a much larger class of matrices.

Dual Certificate Construction: For $\mathbf{M} \in \Gamma^{\perp}$ and $\mathbf{N} \in \mathcal{T}$, consider the infinite sums

$$\begin{aligned}\mathbf{S_M} &= \mathbf{M} - \mathcal{P}_{\mathcal{T}}(\mathbf{M}) + \mathcal{P}_{\Gamma^{\perp}}\mathcal{P}_{\mathcal{T}}(\mathbf{M}) - \mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma^{\perp}}\mathcal{P}_{\mathcal{T}}(\mathbf{M}) + \cdots \\ \mathbf{V_N} &= \mathbf{N} - \mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}) + \mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}) - \mathcal{P}_{\Gamma^{\perp}}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}) + \cdots\end{aligned}$$

Provided that these two sums converge, let $\mathcal{Q} = (1-\eta)\mathbf{V}_{\mathbf{U}\mathbf{U}^T} + \eta\mathbf{S}_{sign(\mathbf{B}^*)}$. It is easy to check that the equality conditions in Lemma 14 are satisfied. It remains to show that (i) the sums converge and (ii) the inequality conditions in Lemma 14 are satisfied. The proof again requires suitable bounds on $\|\mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}))\|_{\infty}$, as well as on $\|\mathcal{P}_{\Gamma^{\perp}}\mathbf{M}\|$, which crucially depend on the

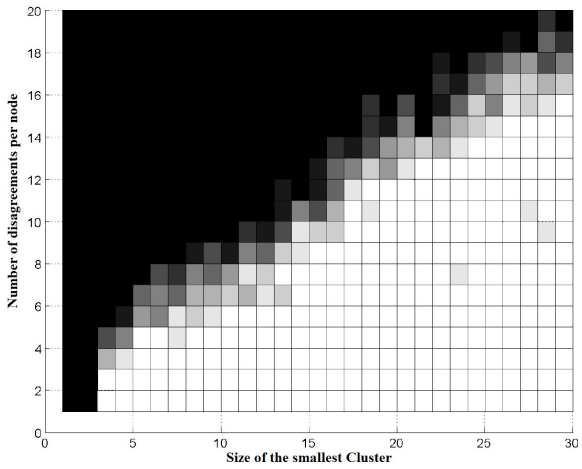


Figure 3.3: Simulation results for fully observed 1000-node graph with cluster of non-uniform sizes. The graph has clusters of at least size k. For different minimum cluster size $K_{\min}$ and number of disgreement per node b, we plot the probability of success.

assumptions imposed on $\mathbf{K}^*$ and $\mathbf{B}^*$; see supplementary materials. Combining the above discussion establishes the theorem.

3.5 Average Case Analysis

In this section, we prove Theorem 7. We first state the probabilistic first-order conditions required for $\mathbf{B}^*$ and $\mathbf{K}^*$ to be the unique optimum of (3.1) with high probability. By *with high probability* we mean with probability at least $1 - cn^{-10}$ for some constant $c > 0$.

Lemma 15 (Probabilistic Sufficient Optimality)**.** *[21] Under the assumptions of Theorem 7,* $\mathbf{K}^*$ *and* $\mathbf{B}^*$ *are unique solutions to* (3.1) *with high proability provided that there exists* $\Omega = \mathbf{W}^B + \mathbf{W}^K$ *such that*

(S1) $\left\|\mathcal{P}_{\mathcal{T}}(\mathbf{W}^B)\right\|_F \leq \frac{1}{2n^2}$.
(S2) $\left\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{W}^B)\right\| < \frac{1}{4}$.
(S3) $\mathcal{P}_{\Omega}(\mathbf{W}^B) = \frac{\eta}{1-\eta} sign(\mathbf{B}^*)$
(S4) $\mathcal{P}_{\Omega_{obs}^\perp}(\mathbf{W}^B) = 0$.
(S5) $\left\|\mathcal{P}_{\Gamma}(\mathbf{W}^B)\right\|_\infty < \frac{1}{4}\frac{\eta}{1-\eta}$.

(L1) $\left\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{W}^K)\right\| < \frac{1}{4}$.
(L2) $\left\|\mathcal{P}_{\mathcal{T}}(\mathbf{W}^K) - \mathbf{U}\mathbf{U}^T\right\|_F \leq \frac{1}{2n^2}$
(L3) $\mathcal{P}_{\Gamma^\perp}(\mathbf{W}^K) = 0$.
(L4) $\left\|\mathcal{P}_{\Gamma}(\mathbf{W}^K)\right\|_\infty < \frac{1}{4}\frac{\eta}{1-\eta}$.

Dual Certificate construction: We used the so-called Golfing Scheme ([21, 58]) to construct $(\mathbf{W}^{\mathrm{B}}, \mathbf{W}^{\mathrm{K}})$. Our application of Golfing Scheme, as well as its analysis, is different from [21], and thus leads to stronger guarantees. In particular, we go beyond existing results by allowing the fraction of observed entries to be vanishing.

With slight abuse of notation, we use Ω_{obs}, Γ, and Ω to denote both the spaces of matrices, as well as the sets of indices these matrices are supported on. By definition, Γ (as a set of indices) contains each entry index with probability $p_0(1-\tau)$. The basic idea is to write Γ as the union of several sets that are independent of each other $\Gamma = \cup_{1\leq k\leq k_0}\Gamma_i$, where each Γ_i contains each entry with probability q, where q and k_0 are suitably chosen. For $1 \leq k \leq k_0$, define the operator $\mathcal{R}_{\Gamma_k}$ by

$$\mathcal{R}_{\Gamma_k}(\mathbf{M}) = \sum_{i=1}^{n} M_{i,i} e_i e_i^T + q^{-1} \sum_{1\leq i<j\leq n} \delta_{ij}^{(k)} M_{i,j} \left(e_i e_j^T + e_j e_i^T\right),$$

where, $\delta_{ij}^{(k)} = 1$ if $(i,j) \in \Gamma_k$ and 0 otherwise, and e_i is the i-th standard basis – i.e., the $n \times 1$ column vector with 1 in its i-th entry and 0 elsewhere. $\mathbf{W}^{\mathrm{B}}$ and $\mathbf{W}^{\mathrm{K}}$ are defined as

$$\mathbf{W}^{\mathrm{B}} = \mathbf{W}_{k_0}^{\mathrm{B}} + \frac{\eta}{1-\eta} sign(\mathbf{B}^*), \quad \mathbf{W}^{\mathrm{K}} = \mathbf{W}_{k_0}^{\mathrm{K}},$$

where, $(\mathbf{W}_{k_0}^{\mathrm{K}}, \mathbf{W}_{k_0}^{\mathrm{B}})$ is defined recursively by setting $\mathbf{W}_0^{\mathrm{B}} = \mathbf{W}_0^{\mathrm{K}} = 0$ and for all $k = 1, 2, \ldots, k_0$,

$$\mathbf{W}_k^{\mathrm{B}} = \mathbf{W}_{k-1}^{\mathrm{B}} - \mathcal{R}_{\Gamma_k}\mathcal{P}_{\mathcal{T}}\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}(sign(\mathbf{B}^*)) + \mathbf{W}_{k-1}^{\mathrm{B}}\right)$$
$$\mathbf{W}_k^{\mathrm{K}} = \mathbf{W}_{k-1}^{\mathrm{K}} + \mathcal{R}_{\Gamma_k}\mathcal{P}_{\mathcal{T}}\left(\mathbf{U}\mathbf{U}^T - \mathbf{W}_{k-1}^{\mathrm{K}}\right).$$

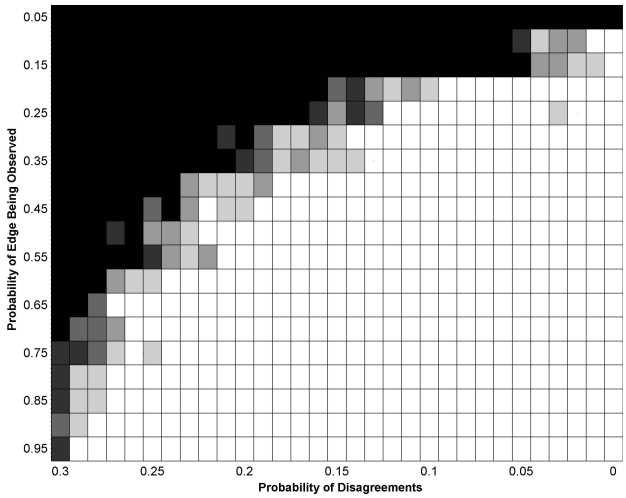


Figure 3.4: Simulation results for partially observed 400-node network with minimum cluster size fixed at $K_{\min} = 60$. Disagreements are placed on each (potential) edge with probability τ, and each edge is observed with probability p_0. The figure shows the probability of success in recovering the ideal cluster under different τ and p_0. Brighter colors show higher success.

It is straightforward to verify that the equality constraints in Lemma 15 are satisfied. Moreover, $\mathbf{W}^{\mathrm{K}}$ satisfies the inequality constraints. The proof is nearly identical to that of Y^L in section 7.3 in [21]. It remains to prove that $\mathbf{W}^{\mathrm{B}}$ also satisfies the corresponding inequalities in Lemma 15. As in the worst case analysis, the proof involves upper-bounding the norms of matrices after certain (random) linear tranformations, such as $\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma_k}\mathcal{P}_{\mathcal{T}}(\mathbf{M})\|$, $\|\mathcal{P}_{\Gamma_k}(\mathbf{M})\|$, $\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma_k}\mathcal{P}_{\mathcal{T}}(\mathbf{M})\|_\infty$, and $\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}(sign(\mathbf{B}^*))\|_\infty$. These bounds are proven again using the assumptions imposed on $\mathbf{B}^*$, $\mathbf{K}^*$, and Ω_{obs}. The details of the proof are given in the Appendix.

3.6 Experimental Results

We explore the performance of our algorithm as a various graph parameters of interest via simulation. We see that the performance matches well with the theory.

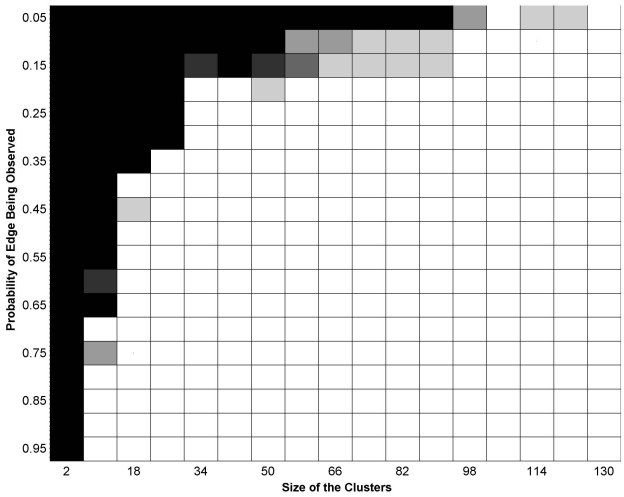


Figure 3.5: Simulation results for partially observed 400-node network with fixed probability $\tau = 0.04$ of placing a disgreement, and different $K_{\min}$ and p_0.

We first verify our deterministic guarantees for fully observed graphs and consider two cases: (1) all clusters have the same size equal to $K_{\min}$, and the number of disagreements involving each node is fixed at b across all nodes; (2) b is again fixed, but clusters may have different sizes no smaller than $K_{\min}$. For each pair $(b, K_{\min})$, a graph is picked randomly from all graphs with the desired property, and we use our algorithm to find $\mathbf{K}^*$ and $\mathbf{B}^*$. The optimization problem (3.1) is solved using the fast algorithm in [84] with η set via line search with step size 0.01. We check if the solution is a valid clustering and is equal to the underlying ideal cluster. The experiment is repeated for 10 times and we plot the probability of success in Fig. 3.2 and 3.3. One can see that the margin of the number of disagreements is higher in the second case, as these graphs have typically larger clusters than in the first case.

We next consider partially observed graphs. A test case is constructed by generating a 400-node graph with equal cluster size $K_{\min}$, and then placing a disagreement on each (potiential) edge with probability τ, independent of all others. Each edge is observed with probability p_0. In the first set of experiments, we fix $K_{\min} = 60$ and vary (p_0, τ). The probability of success is ploted in Fig. 3.5. The second set of experiments have fixed $\tau = 0.04$ and different $(p_0, K_{\min})$, with results ploted in Fig. 3.5. One can see that

our algorithm succeeds with p_0 as small as 10% and the average number of disagreements per node being on the same order of the cluster size. We expect that the fraction of observed entries can be even smaller for larger networks, where the concentration effect is more significant.

3.7 Additional Notations

Definitions related to $\mathbf{K}^*$: For the purpose of analysis only, without loss of generality, by appropriately permuting rows and columns, $\mathbf{K}^*$ can be assumed to be of the block-diagonal form

$$\mathbf{K}^* = \begin{pmatrix} \mathbf{K}_1^* & & & \\ & \mathbf{K}_2^* & & \\ & & . & \\ & & & \mathbf{K}_p^* \end{pmatrix},$$

where each $\mathbf{K}_i^* \in \mathbb{R}^{K_i \times K_i}$ is a matrix with all one entries, and $K_1 \geq K_2 \geq \cdots \geq K_p$, where p is the number of clusters. All other entries of the matrix $\mathbf{K}^*$, i.e., outside these all-one blocks, are zero. We will assume this in all that follows, since all our arguments remain the same if rows and columns are permuted in the same way.

It is easy to show that the SVD of $\mathbf{K}^*$ is of the form $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$, where, $\mathbf{\Sigma} = diag\,(K_1, K_2, \cdots, K_p) \in \mathbb{R}^{p \times p}$ and $\mathbf{U} \in \mathbb{R}^{n \times p}$, where, for all $i \in \{1, 2, \cdots, p\}$, each column $\mathbf{u}_i$ has the following form

$$\mathbf{u}_i = \frac{1}{\sqrt{K_i}} \begin{pmatrix} \mathbf{0}_{\sum_{j=1}^{i-1} K_j} \\ \mathbf{1}_{K_i} \\ \mathbf{0}_{n - \sum_{j=1}^{i} K_j} \end{pmatrix}_{n \times 1}.$$

That is, $\mathbf{u}_i$ is non-zero only in those rows that correspond to nodes in cluster i; these non-zero entries are all equal to $\frac{1}{\sqrt{K_i}}$.

We now define a sub-space $\mathcal{T}$ of the set of all matrices that share either the same column space or the same row space as $\mathbf{K}^*$:

$$\mathcal{T} \;=\; \left\{ \mathbf{U}\mathbf{X}^T + \mathbf{Y}\mathbf{U}^T : \mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times p} \right\}.$$

Now, for an arbitrary matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, we can define its *orthogonal projection* to the space $\mathcal{T}$ as follows:

$$\mathcal{P}_{\mathcal{T}}(\mathbf{M}) = \mathbf{U}\mathbf{U}^T\mathbf{M} + \mathbf{M}\mathbf{U}\mathbf{U}^T - \mathbf{U}\mathbf{U}^T\mathbf{M}\mathbf{U}\mathbf{U}^T.$$

Note that $\mathcal{P}_{\mathcal{T}}(\mathbf{M})$ is also a matrix. We will also be interested in projections onto $\mathcal{T}^{\perp}$, the complement orthogonal space of $\mathcal{T}$ – i.e., the set of all matrices that have zero inner-product with all matrices in $\mathcal{T}$. The projection of any matrix $\mathbf{M}$ onto $\mathcal{T}^{\perp}$ is as follows:

$$\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{M}) \;=\; \mathbf{M} - P_{\mathcal{T}}(\mathbf{M}).$$

Definitions related to $\mathbf{B}^*$: The symmetric matrix $\mathbf{B}^*$ represents the disagreements between the given graph and $\mathbf{K}^*$. We reorder the rows and columns of this as well, in a way that is consistent with the re-ordering of $\mathbf{K}^*$ as described above. Thus the first K_1 rows (and columns) of $\mathbf{B}^*$ correspond to nodes in cluster 1, the next K_2 to nodes in cluster 2, and so on. Thus we have that

$$\mathbf{B}^* = \begin{pmatrix} \mathbf{B}^*_{1,1} & \mathbf{B}^*_{1,2} & \cdot & \mathbf{B}^*_{1,p} \\ \mathbf{B}^{*T}_{1,2} & \mathbf{B}^*_{2,2} & \cdot & \mathbf{B}^*_{2,p} \\ \cdot & \cdot & \cdot & \cdot \\ \mathbf{B}^{*T}_{1,p} & \mathbf{B}^{*T}_{2,p} & \cdot & \mathbf{B}^*_{p,p} \end{pmatrix}.$$

Now, for any two clusters $i, j \in \{1, 2, \cdots, p\}$, the entries of $\mathbf{B}^*_{i,i}$ are either -1, corresponding to the missing edges inside the cluster i, or 0; the entries in $\mathbf{B}^*_{i,j}$ are either $+1$, corresponding to the edges between clusters i and j, or 0.

For any matrix $\mathbf{M}$ define $supp(\mathbf{M}) = \{(i,j) : m_{i,j} \neq 0\}$ to be the support set of the matrix $\mathbf{M}$. Let Ω be the space of matrices with support sets that are a subset of the support set of $\mathbf{B}^*$, i.e.,

$$\Omega = \left\{\mathbf{B} \in \mathbb{R}^{n \times n} : supp(\mathbf{B}) \subseteq supp(\mathbf{B}^*)\right\}.$$

Let $\mathcal{P}_{\Omega}(\mathbf{N}) \in \mathbb{R}^{n \times n}$ be the projection of the matrix $\mathbf{N}$ onto the space Ω, i.e.,

$$\left(\mathcal{P}_{\Omega}(\mathbf{N})\right)_{i,j} = \begin{cases} n_{i,j} & \text{if } b^*_{i,j} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

In words, $\mathcal{P}_\Omega(\mathbf{N})$ is obtained from $\mathbf{N}$ by setting all entries not in the set $supp(\mathbf{B}^*)$ to zero.

Let $\Omega^\perp$ be the orthogonal space to Ω – it is the space of all matrices whose entries in the set $supp(\mathbf{B}^*)$ are zero. The projection onto $\Omega^\perp$ is as follows

$$\mathcal{P}_{\Omega^\perp}(\mathbf{N}) = \mathbf{N} - \mathcal{P}_\Omega(\mathbf{N}).$$

Now, $\mathcal{P}_{\Omega^\perp}(\mathbf{N})$ is obtained from $\mathbf{N}$ by setting all entries *in* the set $supp(\mathbf{B}^*)$ to zero.

Extra definitions for partially observed case: Let Ω_{obs} be the space of matrices with support sets that are a subset of the set of observed entries and

$$\Gamma = \Omega^\perp \cap \Omega_{\text{obs}},$$

is the set of matrices with support within the set of observed entries but outside the set of disgreements. Accordingly, define $\mathcal{P}_{\Omega_{\text{obs}}}$, $\mathcal{P}_{\Omega^\perp_{\text{obs}}}$, $\mathcal{P}_\Gamma$ and $\mathcal{P}_{\Gamma^\perp}$ similar to the definition of $\mathcal{P}_{\Omega^\perp}$ and $\mathcal{P}_\Omega$.

In our probabilistic analysis, we assume that each of possible $\binom{n}{2}$ disagreement edges is present with probability τ. Accordingly, for each present edge, we set $b^*_{i,j}$'s to be $+1$ if the edge is between two clusters and -1 otherwise. Similarly, we assume that each edge is observed with probability p_0. Due to the symmetric structure of $\mathbf{B}^*$, observing an edge is equivalent to observing two (equal) entries of the matrix.

Norms: We now define the several matrix norms we need to use in the following. We use $\|\mathbf{M}\|$ to represent the spectral norm of the matrix $\mathbf{M}$. $\|\mathbf{M}\|_*$ is the nuclear norm of the matrix $\mathbf{M}$ and is equal to the sum of the singular values of the matrix $\mathbf{M}$. With slightly abuse of notation, we extend vector ℓ_1 and ℓ_∞ norms to matrices – we define $\|\mathbf{M}\|_1 = \sum_{i,j} |m_{i,j}|$ to be the sum of the absolute values of all entries of the matrix $\mathbf{M}$ and $\|\mathbf{M}\|_\infty = \max_{i,j} |m_{i,j}|$ to be the element-wise maximum magnitude of the matrix $\mathbf{M}$. We also use $\|\cdot\|_F$ to denote the Frobenius norm.

3.8 Proof of Theorem 6

We prove Theorem 6 in this section. Recall that we need to prove $\mathcal{T} \cap \Gamma^{\perp} = \{\mathbf{0}\}$, and show the existence of a dual certificate $\mathcal{Q}$ obeying the following sufficient optimality conditions:

(a) $\mathcal{P}_{\Omega^{\perp}_{\text{obs}}}(\mathcal{Q}) = 0$

(b) $\mathcal{P}_{\mathcal{T}}(\mathcal{Q}) = (1-\eta)\mathbf{U}\mathbf{U}^T$.

(c) $\mathcal{P}_{\Omega}(\mathcal{Q}) = \eta sign(\mathbf{B}^*)$.

(d) $\|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathcal{Q})\| < 1-\eta$.

(e) $\|\mathcal{P}_{\Omega^{\perp}}(\mathcal{Q})\|_{\infty} < \eta$.

We propose to construct $\mathcal{Q}$ as follows. For $\mathbf{M} \in \Gamma^{\perp}$ and $\mathbf{N} \in \mathcal{T}$, consider the infinite sums

$$\begin{aligned}\mathbf{S_M} &= \mathbf{M} - \mathcal{P}_{\mathcal{T}}(\mathbf{M}) + \mathcal{P}_{\Gamma^{\perp}}\mathcal{P}_{\mathcal{T}}(\mathbf{M}) - \mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma^{\perp}}\mathcal{P}_{\mathcal{T}}(\mathbf{M}) + \cdots \\ \mathbf{V_N} &= \mathbf{N} - \mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}) + \mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}) - \mathcal{P}_{\Gamma^{\perp}}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}) + \cdots\end{aligned}$$

Provided that these two sums converge, we let

$$\mathcal{Q} = (1-\eta)\mathbf{V}_{\mathbf{U}\mathbf{U}^T} + \eta \mathbf{S}_{sign(\mathbf{B}^*)}.$$

The convergence of the infinite sum is guaranteed by Lemma 18 in the next subsection. It is also easy to check that $\mathcal{Q}$ satisfies the equality conditions. The following lemma proves $\mathcal{T} \cap \Gamma^{\perp} = \{\mathbf{0}\}$.

Lemma 16. *Under the assumptions of Theorem 6,* $\mathcal{T} \cap \Gamma^{\perp} = \{\mathbf{0}\}$.

Proof. Using [31] (see Proposition 1 therein and replace Ω with $\Gamma^{\perp}$), it suffices to show that $nD_{\max}\frac{1}{K_{\min}} < 1$, which is implied by our assumption. □

It remains to show that $\mathcal{Q}$ satisfies the inequality conditions. The next lemma provides this result. The proof utilizes the auxiliary lemmas given in the next sub-section. Define $\alpha = 3\left(1 - \frac{1}{K_1}\right)D_{\max} + \frac{1}{K_p^2}$.

Lemma 17. *Under the assumption of Theorem 6, $\mathcal{Q}$ satisfies inequality conditions.*

Proof. W.L.O.G. we only consider the non-degenerate case where there is at least 1 disgreement and 2 clusters in the graph, i.e., $D_{\max} \geq \frac{1}{n}$ and $K_{\min} \leq \frac{n}{2}$. Under the assumption of Theorem 6, we have $\alpha < \frac{1}{4}$ and the range for η is non-empty. By Lemma 19 and sum of the geometric series, we have

$$\begin{aligned}
&\|\mathcal{P}_{\Omega^\perp}(\mathcal{Q})\|_\infty \\
&\leq \frac{1}{1-\alpha}\left\|(1-\eta)\mathcal{P}_\Omega\left(\mathbf{U}\mathbf{U}^T\right) - \eta sign(\mathbf{B}^*)\right\|_\infty \\
&\leq \frac{1}{1-\alpha}\frac{1-\eta}{K_{\min}} + \frac{\alpha}{1-\alpha}\eta \\
&< \eta.
\end{aligned}$$

Here, we used the special structure of $\mathbf{U}\mathbf{U}^T$. In the second inequality, we used triangle inequality to get the result. The strict inequality holds if $\eta > \frac{1}{1+(1-2\alpha)K_{\min}}$. Moreover, by the result of Lemma 20 for the spectral norm of elements of $\Gamma^\perp$, we have

$$\begin{aligned}
&\|\mathcal{P}_{\mathcal{T}^\perp}(\mathcal{Q})\| \\
&\leq \left\|\mathcal{P}_{\Gamma^\perp}\left(\sum_{i=0}^{\infty}\left(\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma^\perp}\right)^n\left((1-\eta)\mathcal{P}_{\Gamma^\perp}\left(\mathbf{U}\mathbf{U}^T\right) + \eta sign(\mathbf{B}^*)\right)\right)\right\| \\
&\leq nD_{\max}\left\|\mathcal{P}_{\Gamma^\perp}\left(\sum_{i=0}^{\infty}\left(\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma^\perp}\right)^n\left((1-\eta)\mathcal{P}_{\Gamma^\perp}\left(\mathbf{U}\mathbf{U}^T\right) + \eta sign(\mathbf{B}^*)\right)\right)\right\|_\infty
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
\|\mathcal{P}_{\mathcal{T}^\perp}(\mathcal{Q})\| &\leq \frac{nD_{\max}}{1-\alpha}\left\|(1-\eta)\mathcal{P}_\Omega(\mathbf{U}\mathbf{U}^T) - \eta sign(\mathbf{B}^*)\right\|_\infty \\
&\leq \frac{nD_{\max}}{1-\alpha}\left((1-\eta)\frac{1}{K_{\min}} + \eta\right) \\
&< 1-\eta.
\end{aligned}$$

The strict inequality holds if $\eta < 1 - \frac{K_{\min}}{\left(1+\frac{1-\alpha}{nD_{\max}}\right)K_{\min}-1}$. Combining the conditions on η, we need

$$\frac{1}{1+(1-2\alpha)K_{\min}} < \eta < 1 - \frac{K_{\min}}{\left(1+\frac{1-\alpha}{nD_{\max}}\right)K_{\min}-1},$$

which are implied by range of η in Theorem 6. □

3.8.1 Auxiliary Lemmas

In this sub-section we provide several lemmas required in the preceding proofs.

Lemma 18. *If* $\mathcal{T} \cap \Gamma^{\perp} = \{\mathbf{0}\}$ *then for any* $\mathbf{M} \in \Gamma^{\perp}$ *and* $\mathbf{N} \in \mathcal{T}$ *the series* $\mathbf{S_M}$ *and* $\mathbf{V_N}$ *converge.*

Proof. The proof follows from the fact that if $\Gamma^{\perp} \cap \mathcal{T} = \{\mathbf{0}\}$, then the projection of an element of one of these spaces into the other is a contraction. More formally, Let $\{\mathbf{T}_1, \mathbf{T}_2, \cdots, \mathbf{T}_a\}$ and $\{\omega_1, \omega_2, \cdots, \omega_b\}$ be orthonormal basis for the spaces T and $\Gamma^{\perp}$, respectively. Let $\mathbf{T}_c$ and ω_c be unit-length combinations of $\{\mathbf{T}_i\}$'s and $\{\omega_i\}$'s, respectively. Since $\Gamma^{\perp} \cap \mathcal{T} = \{\mathbf{0}\}$, there exists $\theta < 1$ such that $\langle \mathbf{T}_c, \omega_c \rangle = \langle \mathbf{T}_c \omega_c, \leq \rangle \theta$ for all $\mathbf{T}_c$ and ω_c. Thus, we have $\|\mathcal{P}_{\Gamma^{\perp}}(\mathcal{P}_{\mathcal{T}}(\mathbf{M}))\|_F \leq \theta^2 \|\mathbf{M}\|_F$ and $\mathbf{S_M}$ converges geometrically fast. With similar argument, one can show that $\mathbf{V_N}$ also converges. □

Lemma 19. *For any* $\mathbf{N} \in \mathcal{T}$*, we have* $\|\mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}))\|_{\infty} \leq \alpha \|\mathbf{N}\|_{\infty}$.

Proof. For any matrix $\mathbf{M} \in \Gamma^{\perp}$, by linearity of the projection, we have $\|\mathcal{P}_{\mathcal{T}}(\mathbf{M})\|_{\infty} \leq \|\mathcal{P}_{\mathcal{T}}(sign(\mathbf{M}))\|_{\infty} \|\mathbf{M}\|_{\infty}$. Denote $\mathcal{P}_{\Gamma^{\perp}}(\mathbf{1})$ by $\mathbf{1}^{\Gamma^{\perp}}$. For any entry (k, l) belonging to the submatrix $\mathbf{B}^*_{i,j}$ (WLOG $K_i \geq K_j$), we have

$$
\begin{aligned}
&\left|(\mathcal{P}_{\mathcal{T}}(sign(\mathbf{M})))_{k,l}\right| \\
&\quad \leq \left(\frac{1}{K_i} - \frac{1}{K_i K_j}\right) \sum_{t=1}^{K_i} \mathbf{1}^{\Gamma^{\perp}}_{t,l} + \left(\frac{1}{K_j} - \frac{1}{K_i K_j}\right) \sum_{q=1}^{K_j} \mathbf{1}^{\Gamma^{\perp}}_{k,q} \\
&\qquad + \frac{1}{K_i K_j} \sum_{k \neq t=1}^{K_i} \sum_{l \neq q=1}^{K_j} \mathbf{1}^{\Gamma^{\perp}}_{t,q} + \frac{1}{K_i K_j} \\
&\quad \leq \left(\frac{1}{K_i} - \frac{1}{K_i K_j}\right) D_{\max} K_j + \left(\frac{1}{K_j} - \frac{1}{K_i K_j}\right) D_{\max} K_j \\
&\qquad + \frac{1}{K_i K_j} D_{\max} K_j (K_j - 1) + \frac{1}{K_i K_j} \\
&\quad \leq 3\left(1 - \frac{1}{K_1}\right) D_{\max} + \frac{1}{K_p^2} = \alpha.
\end{aligned}
$$

This concludes the proof of the lemma. □

Lemma 20. *For any* $\mathbf{M} \in \Gamma^{\perp}$*, we have* $\|\mathbf{M}\| \leq nD_{\max}\|\mathbf{M}\|_{\infty}$.

Proof. Note that $\|\mathbf{M}\| \leq \|\mathbf{M}_{\sigma}\|$, where, $\mathbf{M}_{\sigma} \in \mathbb{R}^{p\times p}$ with $(\mathbf{M}_{\sigma})_{i,j} = \|\mathbf{M}_{i,j}\|$. Moreover, by definition of $d_{i,j}$, we have $\|\mathbf{M}_{i,j}\| \leq d_{i,j}$ and hence, $\|\mathbf{M}_{i,j}\| \leq D_{\max}K_j\|\mathbf{M}_{i,j}\|_{\infty}$, assuming $i \leq j$ without loss of generality. Thus, $\|\mathbf{M}\| \leq D_{\max}\|\mathbf{K}_{\sigma}\|\|\mathbf{M}\|_{\infty}$, where, $\mathbf{K}_{\sigma}$ is called a Parisi matrix and has the form

$$\mathbf{K}_{\sigma} = \begin{pmatrix} K_1 & K_2 & \cdot & K_p \\ K_2 & K_2 & \cdot & K_p \\ \cdot & \cdot & \cdot & \cdot \\ K_p & K_p & \cdot & K_p \end{pmatrix}.$$

It is easy to show that $\|\mathbf{K}_{\sigma}\| \leq \sum_{i=1}^{p} K_i = n$. Hence, the result follows. □

3.9 Proof of Theorem 7

We prove Theorem 7 in this section.

The first observation is that, using similar elimination and derandomization arguments as in [21], it suffices to prove the theorem under random sign assumption of $\mathbf{B}^*$, i.e., the signs of the nonzero upper-triangular entries of $\mathbf{B}^*$ are $+$ or $-$ with equal probabilities (the lower-triangular part is symmetric). We briefly explain the reason. When $\tau < 1/2$, we can think of $\mathbf{B}^*$ as a random-signed matrix $\tilde{\mathbf{B}}^*$ with half of its nonzero entries set to zero. If our algorithm succeeds when the disagreements are given by $\tilde{\mathbf{B}}^*$, it also succeeds when there are fewer disagreements. We refer the reader to Theorem 2.3 in [21] for rigorous proof of this argument.

Recall that we need to show the existence of a dual certificate $\mathcal{Q} = \mathbf{W}^{\mathrm{B}} + \mathbf{W}^{\mathrm{K}}$ obeying the following sufficient optimality conditions:

(S1) $\left\|\mathcal{P}_{\mathcal{T}}(\mathbf{W}^{\mathrm{B}})\right\|_F \leq \frac{1}{2n^2}$. (L1) $\left\|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{W}^{\mathrm{K}})\right\| < \frac{1}{4}$.
(S2) $\left\|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{W}^{\mathrm{B}})\right\| < \frac{1}{4}$. (L2) $\left\|\mathcal{P}_{\mathcal{T}}(\mathbf{W}^{\mathrm{K}}) - \mathbf{U}\mathbf{U}^{T}\right\|_F \leq \frac{1}{2n^2}$.
(S3) $\mathcal{P}_{\Omega}(\mathbf{W}^{\mathrm{B}}) = \frac{\eta}{1-\eta} sign(\mathbf{B}^*)$
(S4) $\mathcal{P}_{\Omega_{\mathrm{obs}}^{\perp}}(\mathbf{W}^{\mathrm{B}}) = 0$. (L3) $\mathcal{P}_{\Gamma^{\perp}}(\mathbf{W}^{\mathrm{K}}) = 0$.
(S5) $\left\|\mathcal{P}_{\Gamma}(\mathbf{W}^{\mathrm{B}})\right\|_{\infty} < \frac{1}{4}\frac{\eta}{1-\eta}$. (L4) $\left\|\mathcal{P}_{\Gamma}(\mathbf{W}^{\mathrm{K}})\right\|_{\infty} < \frac{1}{4}\frac{\eta}{1-\eta}$.

We used the so-called Golfing Scheme ([21, 58]) to construct $(\mathbf{W}^{\mathrm{B}}, \mathbf{W}^{\mathrm{K}})$. With slight abuse of notation, we use Ω_{obs}, Γ, and Ω to denote both the

spaces of matrices, as well as the sets of indices these matrices are supported on. By definition, Ω and Ω_{obs} (as sets of entries) obey $\Omega^c \sim \text{Ber}(1-\tau)$ and $\Omega_{\text{obs}} \sim \text{Ber}(p_0)$. Observe that Ω^c may be considered to be generated by $\cup_{1\leq k\leq k_0}\Omega_k$, where the sets $\Omega_k \sim \text{Ber}(q_1)$ are independent; here the parameter q_1 obeys $1-\tau = 1-(1-q_1)^{k_0}$, and k_0 is chosen to be $\lceil 4\log n\rceil$. Similarly, we think of $\Omega_{\text{obs}} \sim \text{Ber}(p_0)$ as $\bigcup_{1\leq k\leq k_0}\Omega_{\text{obs},k}$, where the sets $\Omega_k \sim \text{Ber}(q_2)$ are independent and q_2 obeys $p_0 = 1-(1-q_2)^{k_0}$. One can verify that Ω and Ω_{obs} generated as above have the same distribution as before. Define $\Gamma_k = \Omega_k \cap \Omega_{\text{obs},k}$; we have $\Gamma_k \sim \text{Ber}(q)$ with $q := q_1q_2 \geq p_0(1-\tau)/k_0^2 \geq C_0\frac{n\log n}{K_{\min}^2}$ for some constant C_0. For any random set of symmetric entries $\Omega_0 \sim \text{Ber}(p)$, define the operator $\mathcal{R}_{\Omega_0}$ by

$$\begin{aligned}&\mathcal{R}_{\Omega_0}(\mathbf{M})\\ =\ &\sum_{i=1}^n m_{i,i}e_ie_i^T + p^{-1}\sum_{1\leq i<j\leq n}\delta_{ij}m_{i,j}\left(e_ie_j^T + e_je_i^T\right),\end{aligned}$$

where $\delta_{ij}=1$ if $(i,j)\in\Omega_0$ and 0 otherwise, and e_i is the i-th standard basis – i.e., the $n\times 1$ column vector with 1 in its i-th entry and 0 elsewhere.

We now define our dual certificate. Let $\mathbf{W}^{\text{B}}$ and $\mathbf{W}^{\text{K}}$ be given by

$$\begin{aligned}\mathbf{W}^{\text{B}} &= \mathbf{W}^{\text{B}}_{k_0} + \frac{\eta}{1-\eta}sign(\mathbf{B}^*)\\ \mathbf{W}^{\text{K}} &= \mathbf{W}^{\text{K}}_{k_0},\end{aligned}$$

where $(\mathbf{W}^{\text{K}}_{k_0}, \mathbf{W}^{\text{B}}_{k_0})$ is defined recursively by setting $\mathbf{W}^{\text{B}}_0 = \mathbf{W}^{\text{K}}_0 = 0$ and for all $k = 1, 2, \ldots, k_0$,

$$\begin{aligned}\mathbf{W}^{\text{B}}_k &= \mathbf{W}^{\text{B}}_{k-1} - \mathcal{R}_{\Gamma_k}\mathcal{P}_{\mathcal{T}}\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}(sign(\mathbf{B}^*)) + \mathbf{W}^{\text{B}}_{k-1}\right)\\ \mathbf{W}^{\text{K}}_k &= \mathbf{W}^{\text{K}}_{k-1} + \mathcal{R}_{\Gamma_k}\mathcal{P}_{\mathcal{T}}\left(\mathbf{U}\mathbf{U}^T - \mathbf{W}^{\text{K}}_{k-1}\right).\end{aligned}$$

It is straightforward to verify that the equality conditions are satisfied. Moreover, $\mathbf{W}^{\text{K}}$ satisfies the inequality conditions – the proof is nearly identical to that of Y^L in section 7.3 in [21]; the only difference is that, in order to accommodate the case of symmetric matrices here, we use the auxiliary lemmas 22, 23 and 24 in the next subsection instead of the asymmetric counterparts

Theorem 2.6, Lemma 3.2 and Lemma 3.1 in [21]. Note that the quantities μr and λ in [21] should be translated to the quantities $\frac{n^2}{K_{\min}^2}$ and $\frac{\eta}{1-\eta}$, respectively, in our setup.

It remains to show that $\mathbf{W}^{\mathrm{B}}$ also satisfies the corresponding inequality conditions (S1), (S2) and (S5) with high probability. The proof make use of the auxiliary lemmas given in the next subsection. For convenience of notation, define the quantity $\Delta_k = -\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}\left(sign(\mathbf{B}^*)\right) - \mathcal{P}_{\mathcal{T}}(\mathbf{W}_k^{\mathrm{B}})$, and write

$$\prod_{i=1}^{k}(\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_i}\mathcal{P}_{\mathcal{T}}) = (\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_k}\mathcal{P}_{\mathcal{T}})\cdots(\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_1}\mathcal{P}_{\mathcal{T}})$$

where the order of multiplication is important. Observe that by construction of $\mathbf{W}^{\mathrm{B}}$, we have

$$\Delta_k = -\prod_{i=1}^{k}(\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_i}\mathcal{P}_{\mathcal{T}})\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}(sign(\mathbf{B}^*)) \tag{3.3}$$

$$\mathbf{W}_{k_0}^{\mathrm{B}} = \sum_{k=1}^{k_0}\mathcal{R}_{\Gamma_k}\Delta_{k-1} \tag{3.4}$$

We will also make use of the following estimate, which follows from the structure of $\mathbf{U}$.

$$\left\|\mathcal{P}_{\mathcal{T}}(e_ie_j^\top)\right\|_F^2 = \|\mathbf{U}\mathbf{U}^Te_i\|^2 + \|\mathbf{U}\mathbf{U}^Te_j\|^2 - \|\mathbf{U}\mathbf{U}^Te_i\|^2\|\mathbf{U}\mathbf{U}^Te_j\|^2 \le \frac{2n}{K_{\min}^2}, \quad \forall 1 \le i,j \le n$$

Inequality (S1): Bounding $\left\|\mathcal{P}_{\mathcal{T}}(\mathbf{W}^{\mathbf{B}})\right\|_F$.

We have the following geometric convergence thanks to (3.3).

$$\begin{aligned}
&\left\|\mathcal{P}_{\mathcal{T}}(\mathbf{W}^{\mathrm{B}})\right\|_F \\
=\ &\left\|\mathcal{P}_{\mathcal{T}}\mathbf{W}_{k_0}^{\mathrm{B}} + \frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right\|_F = \|\Delta_{k_0}\|_F \\
\le\ &\left(\prod_{k=1}^{k_0}\|\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_k}\mathcal{P}_{\mathcal{T}}\|\right)\left\|\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right\|_F \\
\overset{(a)}{\le}\ &\frac{\eta}{1-\eta}e^{-k_0}\left\|\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right\|_F \overset{(b)}{\le} \frac{1}{\sqrt{p_0 n}}n^{-4}\cdot n \\
\overset{(c)}{\le}\ &\frac{1}{\log^2 n}n^{-3} \le \frac{1}{2n^2}.
\end{aligned}$$

Here, (a) uses Lemma 22 with $\epsilon_1 = e^{-1}$, (b) uses our choices of η and k_0 and the fact that $\|\mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*)\|_F \le \|sign(\mathbf{B}^*)\|_F \le n$, and (c) uses the assumption $p_0 \ge C_0^2 \frac{n(\log n)^4}{K_{\min}^2} \ge \frac{(\log n)^4}{n}$.

Inequality (S5): Bounding $\left\|\mathcal{P}_\Gamma(\mathbf{W}^{\mathbf{B}})\right\|_\infty$.

We have

$$\begin{aligned}
&\left\|\mathcal{P}_\Gamma(\mathbf{W}^{\mathrm{B}})\right\|_\infty = \left\|\mathcal{P}_\Gamma\left(\mathbf{W}^{\mathrm{B}}_{k_0} + \frac{\eta}{1-\eta} sign(\mathbf{B}^*)\right)\right\|_\infty \\
= \ & \left\|\mathbf{W}^{\mathrm{B}}_{k_0}\right\|_\infty \overset{(a)}{\le} \sum_{k=1}^{k_0} \left\|\mathcal{R}_{\Gamma_i}\Delta_{i-1}\right\|_\infty \ \le \ q^{-1}\sum_{k=1}^{k_0}\|\Delta_{k-1}\|_\infty \\
\overset{(b)}{=} \ & \sum_{k=1}^{k_0} q^{-1}\left\|\prod_{i=1}^{k-1}(\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_i}\mathcal{P}_{\mathcal{T}})\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}(sign(\mathbf{B}^*))\right\|_\infty
\end{aligned}$$

Here, (a) uses (3.4) and (b) uses (3.3). Consider the k-th term in the summation. We have

$$\begin{aligned}
& q^{-1}\left\|\prod_{i=1}^{k-1}(\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_i}\mathcal{P}_{\mathcal{T}})\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*)\right)\right\|_\infty \\
= \ & \frac{\eta}{1-\eta} q^{-1}\max_{a,b}\left|\left\langle e_a e_b^\top, \ \prod_{i=1}^{k-1}(\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_i}\mathcal{P}_{\mathcal{T}})(\mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*))\right\rangle\right| \\
= \ & \frac{\eta}{1-\eta} q^{-1}\max_{a,b}\left|\left\langle \mathcal{P}_{\Omega_{\mathrm{obs}}\cap\Omega}\prod_{i=1}^{k-1}(\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_{k-i}}\mathcal{P}_{\mathcal{T}})\left(e_a e_b^\top\right), \ sign(\mathbf{B}^*)\right\rangle\right|;
\end{aligned}$$

here in the last equality we use the self-adjointness of the operators. Conditioned on Ω_{obs}, Ω, and Γ_i's, $sign(\mathbf{B}^*)$ has i.i.d. symmetric ± 1 entries, so

Hoeffding's inequality gives,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\eta}{1-\eta}q^{-1}\left|\left\langle \mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}\prod_{i=1}^{k-1}(\mathcal{P}_{\mathcal{T}}-\mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_{k-i}}\mathcal{P}_{\mathcal{T}})\left(e_ae_b^\top\right),\ sign(\mathbf{B}^*)\right\rangle\right| > t|\Omega_{\text{obs}},\Omega,\Gamma_i\text{'s}\right)\\
\leq\ & 2\exp\left(-\frac{2t^2}{\left\|\frac{\eta}{1-\eta}q^{-1}\mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}\prod_{i=1}^{k-1}(\mathcal{P}_{\mathcal{T}}-\mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_{k-i}}\mathcal{P}_{\mathcal{T}})\left(e_ae_b^\top\right)\right\|_F^2}\right)\\
\leq\ & 2\exp\left(-\frac{2t^2}{\left(\frac{\eta}{1-\eta}\right)^2q^{-2}\left\|\mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}\prod_{i=1}^{k-1}(\mathcal{P}_{\mathcal{T}}-\mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_{k-i}}\mathcal{P}_{\mathcal{T}})\right\|^2\|\mathcal{P}_{\mathcal{T}}(e_ae_b)\|_F^2}\right)\\
\leq\ & 2\exp\left(-\frac{t^2}{\left(\frac{\eta}{1-\eta}\right)^2q^{-2}\prod_{i=1}^{k-1}\left\|\mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}(\mathcal{P}_{\mathcal{T}}-\mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_{k-i}}\mathcal{P}_{\mathcal{T}})\right\|^2\frac{2n}{K_{\min}^2}}\right);
\end{aligned}\tag{3.5}$$

here the last inequality uses $\left\|\mathcal{P}_{\mathcal{T}}(e_ae_b^\top)\right\|_F^2\leq\frac{\mu r}{n}$. Under the event

$$G_k:=\left\{\mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}\leq\sqrt{2p_0},\left\|\mathcal{P}_{\mathcal{T}}-\mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_{k-i}}\mathcal{P}_{\mathcal{T}}\right\|\leq\frac{1}{2},\ i=1,\dots k-1\right\},$$

we can integrate out the conditions in (3.5) and obtain

$$\begin{aligned}
&\mathbb{P}\left(\frac{\eta}{1-\eta}q^{-1}\left|\left\langle\prod_{i=1}^{k-1}(\mathcal{P}_{\mathcal{T}}-\mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_{k-i}}\mathcal{P}_{\mathcal{T}})\left(e_ae_b^\top\right),\ sign(\mathbf{B}^*)\right\rangle\right| > t|G_k\right)\\
\leq\ & 2\exp\left(-\frac{t^2}{\left(\frac{\eta}{1-\eta}\right)^2q^{-2}\cdot 2p_0\left(\frac{1}{2}\right)^{2k-2}\frac{2n}{K_{\min}^2}}\right)
\end{aligned}$$

Since $\Omega_{\text{obs}}\cap\Omega\subseteq\Omega_{\text{obs}}$, we apply Lemma 22 with $\Omega_0=\Omega_{\text{obs}}$ and $\epsilon_1=\frac{1}{2}$ to

obtain w.h.p.

$$\begin{aligned}\|\mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}\mathcal{P}_{\mathcal{T}}\| &\leq \|\mathcal{P}_{\Omega_{\text{obs}}}\mathcal{P}_{\mathcal{T}}\| = \sqrt{\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_{\text{obs}}}\mathcal{P}_{\mathcal{T}}\|}. \\ &= \sqrt{p_0\left\|\frac{1}{p_0}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_{\text{obs}}}\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}} + \mathcal{P}_{\mathcal{T}}\right\|} \\ &\leq \sqrt{p_0\left\|\frac{1}{p_0}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_{\text{obs}}}\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\right\| + p_0} \\ &\leq \sqrt{2p_0} \qquad (3.6)\end{aligned}$$

By (3.6) and Lemma 22, we know that the event G_k holds with high probability. Choosing $t = C\left(\frac{1}{2}\right)^{k-1}\frac{\frac{\eta}{1-\eta}}{q}\sqrt{\frac{p_0 n}{K_{\min}^2}\log n}$ with C sufficiently large and using union bound (there is only polynomially many different (a,b)), we conclude that

$$q^{-1}\left\|\prod_{i=1}^{k-1}(\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_i}\mathcal{P}_{\mathcal{T}})\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right)\right\|_\infty \leq C\left(\frac{1}{2}\right)^{k-1}\frac{\frac{\eta}{1-\eta}}{q}\sqrt{\frac{p_0 n}{K_{\min}^2}\log n} \leq \left(\frac{1}{2}\right)^k \cdot \frac{1}{4}\frac{\eta}{1-\eta}$$

with high probability; here the second inequality holds because $q \geq \frac{p_0(1-\tau)}{4\log^2 n}$ by our choice and $p_0 \geq C'\frac{n\log^6 n}{K_{\min}^2}$ by assumption of Theorem 7. Summing over k, it follows that

$$\sum_{k=1}^{k_0} q^{-1}\left\|\prod_{i=1}^{k-1}(\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_i}\mathcal{P}_{\mathcal{T}})\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right)\right\|_\infty \leq \frac{1}{4}\frac{\eta}{1-\eta}, \qquad (3.7)$$

which proves inequality (S1).

Inequality (S2): Bounding $\left\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{W}^{\mathbf{B}})\right\|$.

Observe that by triangle inequality,

$$\left\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{W}^{\text{B}})\right\| \leq \frac{\eta}{1-\eta}\left\|\mathcal{P}_{\mathcal{T}^\perp}(sign(\mathbf{B}^*))\right\| + \left\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{W}^{\text{B}}_{k_0})\right\|.$$

For the first term, it follows from $\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{M})\| \leq \|\mathbf{M}\|$ and a standard argument about the norm of a matrix with i.i.d. entries (e.g., see [140]) that

$\frac{\eta}{1-\eta}\left\|\mathcal{P}_{\mathcal{T}^\perp}(sign(\mathbf{B}^*))\right\| \le 4\sqrt{\tau} \le \frac{1}{8}$ provided τ is sufficiently small. It remains to show that the second term is bounded by $\frac{1}{8}$. To this end, we observe

$$\begin{aligned}
&\left\|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{W}^{\mathrm{B}}_{k_0})\right\| \\
\overset{(a)}{=}\ &\sum_{k=1}^{k_0}\left\|\mathcal{P}_{\mathcal{T}^\perp}\left(\mathcal{R}_{\Gamma_k}\Delta_{k-1}-\Delta_{k-1}\right)\right\| \le \sum_{k=1}^{k_0}\left\|\left(\mathcal{R}_{\Gamma_k}-\mathcal{I}\right)\Delta_{k-1}\right\| \\
\overset{(b)}{=}\ &\sum_{k=1}^{k_0}\left\|\left(\mathcal{R}_{\Gamma_k}-\mathcal{I}\right)\prod_{i=1}^{k-1}\left(\mathcal{P}_{\mathcal{T}}-\mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_i}\mathcal{P}_{\mathcal{T}}\right)\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right)\right\| \qquad (3.8)
\end{aligned}$$

where (a) uses (3.4) and the fact that $\Delta_k \in \mathcal{T}$, and (b) uses (3.3).

The main obstacle in bounding the above expression is that $sign(\mathbf{B}^*)$ and Γ_i's are not independent. To get around this, the key idea is to observe that Γ_i's and $sign(\mathbf{B}^*)$ are independent conditioned on Ω. This is because $\mathrm{supp}(sign(\mathbf{B}^*)) \subseteq \Omega$ is a random subset of the disagreement entries while $\Gamma_i \subseteq \Omega^c$ are random subsets of the non-disagreement entries. To utilized this independence, we decompose the operators in the above equation as a telescoping sum. In particular, if we define the operators

$$\begin{aligned}
\mathcal{A}_k &= \mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Omega_k}\mathcal{P}_{\mathcal{T}} \\
\mathcal{S}_k &= \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Omega_k}\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_k}\mathcal{P}_{\mathcal{T}} \\
\mathcal{B}_k &= \mathcal{R}_{\Omega_k} - \mathcal{I} \\
\mathcal{T}_k &= \mathcal{R}_{\Gamma_k} - \mathcal{R}_{\Omega_k}
\end{aligned}$$

for $k = 1, \ldots, k_0$, then $\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Gamma_k}\mathcal{P}_{\mathcal{T}} = \mathcal{A}_k + \mathcal{S}_k$ and $\mathcal{R}_{\Gamma_k} - \mathcal{I} = \mathcal{B}_k + \mathcal{T}_k$. The reason for doing so is that, conditioned on Ω, $\mathcal{T}_k$'s and $\mathcal{S}_i$'s are independent of $sign(\mathbf{B}^*)$. Thus if a term only involves $\mathcal{T}_k$ and $\mathcal{S}_k$'s (we call it a Type-1 term), it can be bounded using Lemma 23 and 24. For the other terms that involve not only $\mathcal{T}_k$ and $\mathcal{S}_k$'s but also $\mathcal{A}_i$'s and/or $\mathcal{B}_k$'s (dubbed Type-2 terms), we bound them by utilizing the random signs of $sign(\mathbf{B}^*)$.

The details are provided below. Consider the k-th term in summands

of the second term in (3.8). Using the above definitions, we have

$$\begin{aligned} & \left\| (\mathcal{R}_{\Gamma_k} - \mathcal{I}) \prod_{i=1}^{k-1} (\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}} \mathcal{R}_{\Gamma_i} \mathcal{P}_{\mathcal{T}}) \left(-\frac{\eta}{1-\eta} \mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*) \right) \right\| \\ = & \left\| (\mathcal{B}_k + \mathcal{T}_k) \prod_{i=1}^{k-1} (\mathcal{A}_i + \mathcal{S}_i) \left(\frac{\eta}{1-\eta} \mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*) \right) \right\| \end{aligned} \tag{3.9}$$

We expand the product and sums in the above equation, which results in a summmation of 2^k=poly(n) terms since $k \leq k_0 = O(\log n)$. Among them there is one Type-1 term

$$\mathcal{T}_k \mathcal{S}_1 \mathcal{S}_2 \cdots \mathcal{S}_{k-1} \left(\frac{\eta}{1-\eta} \mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*) \right), \tag{3.10}$$

and $2^k - 1$ Type-2 terms, such as

$$\begin{aligned} & \mathcal{T}_k \mathcal{A}_1 \mathcal{S}_2 \mathcal{S}_3 \cdots \mathcal{A}_{k-2} \mathcal{S}_{k-1} \left(\tfrac{\eta}{1-\eta} \mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*) \right), \\ & \mathcal{B}_k \mathcal{S}_1 \mathcal{A}_2 \mathcal{S}_3 \cdots \mathcal{S}_{k-2} \mathcal{A}_{k-1} \left(\tfrac{\eta}{1-\eta} \mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*) \right). \end{aligned}$$

We first bound the Type-1 term (3.10). Conditioned on Ω, we have

$$\begin{aligned} & \left\| \mathcal{T}_k \mathcal{S}_1 \mathcal{S}_2 \cdots \mathcal{S}_{k-1} \left(\frac{\eta}{1-\eta} \mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*) \right) \right\| \\ = & \left\| \left(\frac{1}{q_1 q_2} \mathcal{P}_{\Omega_{\text{obs},k} \cap \Omega_k} - \frac{1}{q_1} \mathcal{P}_{\Omega_k} \right) \prod_{i=1}^{k-1} \left(\frac{1}{q_1 q_2} \mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega_{\text{obs},i} \cap \Omega_i} \mathcal{P}_{\mathcal{T}} - \frac{1}{q_1} \mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega_i} \mathcal{P}_{\mathcal{T}} \right) \left(\frac{\eta}{1-\eta} \mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*) \right) \right\| \\ \overset{(i)}{\leq} & C \left(\frac{1}{q_1} \sqrt{\frac{n \log n}{q_2}} \right) \left(\frac{1}{2} \right)^{k-1} \left\| \frac{\eta}{1-\eta} \mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*) \right\|_\infty \\ \overset{(ii)}{\leq} & C' \sqrt{\frac{n \log n}{q_2 q_1^2}} \left(\frac{1}{2} \right)^k \frac{\eta}{1-\eta} \frac{p_0}{\log n} \\ \overset{(iii)}{\leq} & \frac{1}{16} \left(\frac{1}{2} \right)^k; \end{aligned} \tag{3.11}$$

here in (i) we apply Lemma 23 with $\Omega_0 = \Omega_{\text{obs},k}$ and $\Gamma_0 = \Omega_k$, as well as Lemma 24 with $\Omega_0 = \Omega_{\text{obs},i}$, $\Gamma_0 = \Omega_i$ and $\epsilon_3 = \frac{1}{2} q_1$, (ii) uses Lemma 25, and (iii) holds under the assumption of Theorem 7.

We next bound the remaining (2^k-1) terms of Type 2. To this end, we four collect five useful inequalities. Because $\Omega_i \sim Ber(q_1)$, Lemma 22 with $\Omega_0 = \Omega_i$ and $\epsilon_1 = C\sqrt{\frac{n\log n}{K_{\min}^2}}$ gives that w.h.p.

$$\begin{aligned} \|\mathcal{A}_i\| &= \|\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Omega_i}\mathcal{P}_{\mathcal{T}}\| \\ &\leq C\sqrt{\frac{n\log n}{K_{\min}^2}} \leq C'\sqrt{\frac{p_0}{\log^4 n}} \end{aligned} \tag{3.12}$$

Lemma 22 with $\Omega_0 = \Omega_k$ and $\epsilon_1 = \frac{1}{2}$ shows that w.h.p.

$$\begin{aligned} &\|\mathcal{P}_{\mathcal{T}}\mathcal{B}_k\| = \left\|\frac{1}{q_1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_k} - \mathcal{P}_{\mathcal{T}}\right\| \\ \leq\ & \frac{1}{q_1}\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_k}\| + \|\mathcal{P}_{\mathcal{T}}\| = \frac{1}{q_1}\sqrt{\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_k}\mathcal{P}_{\mathcal{T}}\|} + 1 \\ \leq\ & \frac{1}{q_1}\sqrt{q_1\left\|\frac{1}{q_1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_k}\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\right\| + q_1\|\mathcal{P}_{\mathcal{T}}\|} + 1 \leq C\sqrt{\frac{1}{q_1}} \leq C'\sqrt{\log n} \end{aligned} \tag{3.13}$$

Similarly, we have w.h.p.

$$\begin{aligned} &\|\mathcal{P}_{\mathcal{T}}\mathcal{T}_k\| = \left\|\frac{1}{q_1q_2}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_{\mathrm{obs},k}\cap\Omega_k} - \frac{1}{q_1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_k}\right\| \\ \leq\ & \left\|\frac{1}{q_1q_2}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_{\mathrm{obs},k}\cap\Omega_k} - \mathcal{P}_{\mathcal{T}}\right\| + \left\|\mathcal{P}_{\mathcal{T}} - \frac{1}{q_1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_k}\right\| \\ \leq\ & C\sqrt{\frac{1}{q_1q_2}} + C\sqrt{\frac{1}{q_1}} \leq C'\sqrt{\frac{\log^2 n}{p_0}}. \end{aligned} \tag{3.14}$$

Applying Lemma 22 twice with (1) $\Omega_0 = \Omega_k$, $\epsilon_1 = C\sqrt{\frac{n\log n}{K_{\min}^2 q_1}}$ and (2) $\Omega_0 = \Omega_{\mathrm{obs},k}\cap\Omega_k$, $\epsilon_1 = C\sqrt{\frac{n\log n}{K_{\min}^2 q_1 q_2}}$ gives w.h.p.

$$\begin{aligned} &\|\mathcal{S}_k\| = \left\|\frac{1}{q_1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_k}\mathcal{P}_{\mathcal{T}} - \frac{1}{q_1q_2}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_{\mathrm{obs},k}\cap(\Omega_k)}\mathcal{P}_{\mathcal{T}}\right\| \\ \leq\ & \left\|\frac{1}{q_1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_k}\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\right\| + \left\|\mathcal{P}_{\mathcal{T}} - \frac{1}{q_1q_2}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{(\Omega_{\mathrm{obs},k}\cap\Omega_k)}\mathcal{P}_{\mathcal{T}}\right\| \\ \leq\ & C\sqrt{\frac{n\log n}{K_{\min}^2 q_1}} + C\sqrt{\frac{n\log n}{K_{\min}^2 q_1 q_2}} \leq C'\sqrt{\frac{n\log^3 n}{K_{\min}^2 p_0}} \leq \frac{1}{4\log n} \end{aligned} \tag{3.15}$$

Now consider one of the Type-2 terms

$$\mathcal{X}\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right) \triangleq \mathcal{T}_k\mathcal{S}_1\mathcal{S}_2\cdots\mathcal{S}_{k-2}\mathcal{A}_{k-1}\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right).$$

Let $\mathcal{X}^*$ be the adjoint of $\mathcal{X}$. The last four inequalities (3.12)-(3.15) together with (3.6) with yield w.h.p.

$$\begin{aligned}\|\mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}\mathcal{P}_{\mathcal{T}}\mathcal{X}^*\| &= \|\mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}\mathcal{P}_{\mathcal{T}}\mathcal{A}_{k-1}\mathcal{S}_{k-2}\cdots\mathcal{S}_1\mathcal{P}_{\mathcal{T}}\mathcal{T}_k\| \\ &\leq C\sqrt{p_0}\sqrt{\frac{p_0}{\log^3 n}}\left(\frac{1}{4}\right)^{k-2}\sqrt{\frac{\log^2 n}{p_0}} \\ &\leq C'\sqrt{p_0}\left(\frac{1}{4}\right)^k. \end{aligned} \tag{3.16}$$

It is not hard to check that this inequality also holds for the $\mathcal{X}$'s associated with other Type-2 terms . We are ready to bound the operator norm of the Type-2 term using a standard ϵ-net argument. Let $\mathbb{S}^{n-1}$ be the unit sphere in $\mathbb{R}^n$, and N be an 1/2-net of $\mathbb{S}^{n-1}$ of size at most 6^n. The definition of the net and Lipschitz property of the operator norm gives that

$$\begin{aligned}&\left\|\mathcal{X}\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right)\right\| \\ =\ &\sup_{x,y\in\mathbb{S}^{n-1}}\left\langle xy^\top,\ \mathcal{X}\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right)\right\rangle \\ \leq\ &4\sup_{x,y\in N}\left\langle xy^\top,\ \mathcal{X}\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right)\right\rangle\end{aligned}$$

For a fixed pair $(x,y)\in N\times N$, we have

$$\begin{aligned}&\left\langle xy^\top,\ \mathcal{X}\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*)\right)\right\rangle \\ =\ &\frac{\eta}{1-\eta}\left\langle\mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}\mathcal{P}_{\mathcal{T}}\mathcal{X}^*\left(xy^\top\right),\ sign(\mathbf{B}^*)\right\rangle\end{aligned}$$

We condition on the event that (3.16) holds. Because $sign(\mathbf{B}^*)$ has i.i.d.

symmetric ± 1 entries, Hoeffding's inequality gives

$$\begin{aligned}
& \mathbb{P}\left(\frac{\eta}{1-\eta}\left\langle \mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}\mathcal{P}_{\mathcal{T}}\mathcal{X}^*\left(xy^\top\right),\ sign(\mathbf{B}^*)\right\rangle \geq \frac{C}{4^k}\right) \\
\leq\ & 2\exp\left(-\frac{2\cdot\frac{C^2}{4^{2k}}}{\left\|\frac{\eta}{1-\eta}\mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}\mathcal{P}_{\mathcal{T}}\mathcal{X}^*\left(xy^\top\right)\right\|_F^2}\right) \\
\leq\ & 2\exp\left(-\frac{2\cdot\frac{C^2}{4^{2k}}}{\frac{1}{32^2p_0n}\left\|\mathcal{P}_{\Omega_{\text{obs}}\cap\Omega}\mathcal{P}_{\mathcal{T}}\mathcal{X}^*\right\|^2}\right) \\
\leq\ & 2\exp\left(-\frac{C'\cdot\frac{1}{4^{2k}}}{\frac{1}{np_0}\cdot p_0\frac{1}{4^{2k}}}\right) \\
\leq\ & 2\exp\left(-C'n\right)
\end{aligned}$$

for some constant C' that can be made large. This probability is exponentially small, so we can apply union bound over the 6^n pairs (x, y) in the ϵ-net $N \times N$ and conclude that w.h.p.

$$\left\|\mathcal{X}(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}sign(\mathbf{B}^*))\right\| \leq \frac{C}{4^k} = C\frac{1}{2^k}\frac{1}{2^k}$$

Summing over all $2^k - 1$ Type-2 terms and combining with the bound (3.11) for the Type-1 term, it follows that the right hand side of (3.9) is bounded by $\frac{1}{8\cdot 2^k}$. Summing over $k = 1, 2, \ldots, k_0$ bounds the second term in (3.8) by $\frac{1}{8}$. This completes the proof of inequality (S2) and thus the proof of Theorem 7.

3.9.1 Auxiliary Lemmas

In this sub-section, we provide several auxiliary lemmas required in the above proof. We will make use of the non-commutative Bernstein inequality. The version given below is first proved in [58, 114] and later sharpened in [134].

Lemma 21. *[134] Consider a finite sequence $\{\mathbf{M}_i\}$ of independent, random $n_1 \times n_2$ matrices that satisfy the assumption $\mathbb{E}\mathbf{M}_i = 0$ and $\|\mathbf{M}_i\| \leq D$ almost*

surely. Let

$$\sigma^2 = \max\left\{\left\|\sum_i \mathbb{E}\left[\mathbf{M}_i\mathbf{M}_i^\top\right]\right\|, \left\|\sum_i \mathbb{E}\left[\mathbf{M}_i^\top\mathbf{M}_i\right]\right\|\right\}.$$

Then for all $t > 0$ we have

$$\begin{aligned}\mathbb{P}\left[\left\|\sum \mathbf{M}_i\right\| \geq t\right] &\leq (n_1+n_2)\exp\left(-\frac{t^2}{2\sigma^2 + 2Dt/3}\right). \\ &\leq \begin{cases}(n_1+n_2)\exp\left(-\frac{3t^2}{8\sigma^2}\right), & \text{for } t \leq \frac{\sigma^2}{D}; \\ (n_1+n_2)\exp\left(-\frac{3t}{8D}\right), & \text{for } t \geq \frac{\sigma^2}{D}.\end{cases}\end{aligned} \tag{3.17}$$

Remark 1. *When $n_1 = n_2 = 1$, this becomes the standard two-sided Bernstein inequality.*

The first auxiliary lemma is similar to Theorem 4.1 in [21], but adapted to the symmetric case. Our proof is different from [21].

Lemma 22. *Suppose Ω_0 is a set of entries obeying $\Omega_0 \sim Ber(p)$. Consider the operator $P_{\mathcal{T}} - P_{\mathcal{T}}\mathcal{R}_{\Omega_0}P_{\mathcal{T}}$ restricted on the space of symmetric matrices. For some constant $C_0 > 0$, we have*

$$\|\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Omega_0}\mathcal{P}_{\mathcal{T}}\| < \epsilon_1,$$

with high probability provided that $p \geq C_0\frac{n\log n}{\epsilon_1^2 K_{\min}^2}$ and $\epsilon_1 \leq 1$.

Proof. For each (i,j), define the indicator random variable $\delta_{ij} = \mathbf{1}_{\{(i,j)\in\Omega_0\}}$. We observe that for any matrix $\mathbf{M} \in \mathcal{T}$

$$\begin{aligned}&(\mathcal{P}_{\mathcal{T}}\mathcal{R}_{\Omega_0}\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}})\,\mathbf{M} \\ =\ &\sum_{1\leq i<j\leq n}\left(p^{-1}\delta_{ij} - 1\right)\left\langle \mathcal{P}_{\mathcal{T}}(e_ie_j^\top),\ \mathbf{M}\right\rangle \mathcal{P}_{\mathcal{T}}(e_ie_j^\top + e_je_i^\top) \\ \triangleq\ &\sum_{1\leq i<j\leq n}\mathcal{S}_{ij}(\mathbf{M}).\end{aligned}$$

Here $\mathcal{S}_{ij}:\ \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ is a linear self-adjoint operator with $\mathbb{E}\left[\mathcal{S}_{ij}\right] = 0$. We also have the bounds

$$\begin{aligned}\|\mathcal{S}_{ij}\| &\leq p^{-1}\left\|\mathcal{P}_{\mathcal{T}}(e_ie_j^\top)\right\|_F\left\|\mathcal{P}_{\mathcal{T}}(e_ie_j^\top + e_je_i^\top)\right\|_F \\ &\leq p^{-1}\cdot 2\left\|\mathcal{P}_{\mathcal{T}}(e_ie_j^\top)\right\|_F^2 \leq \frac{4n}{K_{\min}^2 p},\end{aligned}$$

and

$$
\begin{aligned}
&\left\| \mathbb{E}\left[\sum_{1\le i<j\le n} \mathcal{S}_{ij}^2(\mathbf{M}) \right] \right\|_F \\
= &\left\| \sum_{1\le i<j\le n} \mathbb{E}\left[(p^{-1}\delta_{ij}^{(k)} - 1)^2 \right] \left\langle \mathcal{P}_{\mathcal{T}}(e_i e_j^\top),\, \mathbf{M} \right\rangle \left\langle \mathcal{P}_{\mathcal{T}}(e_i e_j^\top + e_j e_i^\top),\, e_i e_j^\top \right\rangle \mathcal{P}_{\mathcal{T}}(e_i e_j^\top + e_j e_i^\top) \right\|_F \\
= &\left(p^{-1} - 1\right) \left\| \sum_{1\le i<j\le n} 2 \left\| \mathcal{P}_{\mathcal{T}}(e_i e_j^\top) \right\|_F^2 m_{i,j} \mathcal{P}_{\mathcal{T}}(e_i e_j^\top + e_j e_i^\top) \right\|_F \\
\le &\left(p^{-1} - 1\right) \left\| \sum_{1\le i<j\le n} 2 \left\| \mathcal{P}_{\mathcal{T}}(e_i e_j^\top) \right\|_F^2 m_{i,j} (e_i e_j^\top + e_j e_i^\top) \right\|_F \\
\le &\left(p^{-1} - 1\right) \frac{4n}{K_{\min}^2} \left\| \sum_{1\le i<j\le n} m_{i,j} (e_i e_j^\top + e_j e_i^\top) \right\|_F \\
= &\left(p^{-1} - 1\right) \frac{4n}{K_{\min}^2} \|\mathbf{M}\|_F ,
\end{aligned}
$$

which means $\left\| \mathbb{E}\left[\sum_{1\le i<j\le n} \mathcal{S}_{ij}^2 \right] \right\| \le \frac{4n}{K_{\min}^2 p}$; here we use the fact that $\mathcal{P}_{\mathcal{T}}(e_i e_j^\top) = \left(\mathcal{P}_{\mathcal{T}}(e_j e_i^\top)\right)^\top$ and $\mathbf{M}$ is symmetric. An application of the Bernstein inequality (first inequality of (3.17)) then yields

$$
\mathbb{P}\left[\left\| \sum_{1\le i<j\le n} \mathcal{S}_{ij} \right\| \ge \epsilon_1 \right] \le 2n^{2-2\beta}
$$

provided $p \ge \frac{64\beta n \log n}{3K_{\min}^2 \epsilon_1^2}$ and $\epsilon_1 < 1$. □

The next lemma is similar to Theorem 6.3 in [22] but adapted to the symmetric case. The proof is again different.

Lemma 23. *Suppose Ω_0 is a set of entries obeying $\Omega_0 \sim Ber(p)$, Γ_0 is a fixed set of pairs of symmetric entries, and $\mathbf{M}$ is a fixed $n \times n$ symmetric matrix. Then for some constant $C_0 > 0$, we have*

$$
\|(\mathcal{P}_{\Gamma_0} - \frac{1}{p}\mathcal{P}_{\Omega_0 \cap \Gamma_0})\mathbf{M}\| < \sqrt{C_0 \frac{n \log n}{p}} \|\mathbf{M}\|_\infty,
$$

with high probability provided that $p \geq C_0 \frac{\log n}{n}$.

Proof. Define δ_{ij} as before. Notice that

$$\frac{1}{p}\mathcal{P}_{\Omega_0 \cap \Gamma_0}(\mathbf{M}) - \mathcal{P}_{\Gamma_0}(\mathbf{M}) = \sum_{i<j,(i,j)\in\Gamma_0} (q^{-1}\delta_{ij} - 1) m_{i,j} \left(e_i e_j^\top + e_j e_i^\top\right) \triangleq \sum_{i<j,(i,j)\in\Gamma_0} S_{ij}.$$

Here the symmetric matrix $S_{ij} \in \mathbb{R}^{n\times n}$ satisfies $\mathbb{E}[S_{ij}] = 0$, $\|S_{ij}\| \leq 2p^{-1}\|\mathbf{M}\|_\infty$ and

$$\begin{aligned}
\left\| \mathbb{E}\left[\sum_{i<j,(i,j)\in\Gamma_0} S_{ij}^2 \right] \right\| &= \left(p^{-1} - 1\right) \left\| \sum_{i<j,(i,j)\in\Gamma_0} m_{i,j}^2 \left(e_i e_i^\top + e_j e_j^\top\right) \right\| \\
&\leq \left(p^{-1} - 1\right) \left\| \mathrm{diag}\left(\sum_{(1,j)\in\Gamma_0} m_{1,j}^2, \ldots, \sum_{(n,j)\in\Gamma_0} m_{n,j}^2 \right) \right\| \\
&\leq \left(p^{-1} - 1\right) n \|\mathbf{M}\|_\infty^2 \leq 2p^{-1} n \|\mathbf{M}\|_\infty^2 .
\end{aligned}$$

When $p \geq \frac{16\beta \log n}{3n}$, we apply the Bernstein inequality (first inequality of (3.17)) and obtain

$$\begin{aligned}
\mathbb{P}\left[\left\| \sum_{i<j,(i,j)\in\Gamma_0} S_{ij} \right\| \geq \sqrt{\frac{16\beta n \log n}{3p}} \|\mathbf{M}\|_\infty \right] &\leq 2n \exp\left(- \frac{3 \cdot \frac{16\beta n \log n}{3p} \|\mathbf{M}\|_\infty^2}{8 \cdot \frac{2n}{p} \|\mathbf{M}\|_\infty^2} \right) \\
&\leq 2n^{1-\beta}.
\end{aligned}$$

The conclusion follows by choosing $\beta > 1$. □

The third lemma is similar to Lemma 3.1 in [21], but extended to the symmetric case. The proof is nearly identical to that in [21].

Lemma 24. *Suppose* Ω_0 *is a set of entries obeying* $\Omega_0 \sim Ber(p)$, Γ_0 *is a fixed set of pairs of symmetric entries, and* $\mathbf{M} \in \mathcal{T}$ *is a fixed symmetric* $n \times n$ *matrix. Then for some constant* $C_0 > 0$, *we have*

$$\|(\mathcal{P}_\mathcal{T}\mathcal{P}_{\Gamma_0}\mathcal{P}_\mathcal{T} - \frac{1}{p}\mathcal{P}_\mathcal{T}\mathcal{P}_{\Omega_0\cap\Gamma_0}\mathcal{P}_\mathcal{T})\mathbf{M}\|_\infty < \epsilon_3 \|\mathbf{M}\|_\infty,$$

with high probability, provided that $p \geq C_0 \frac{n \log n}{\epsilon_3^2 K_{\min}^2}$ *and* $\epsilon_3 \leq 1$.

Proof. Define δ_{ij} as before. Fix an entry index (a,b). Notice that

$$\begin{aligned}\left(\frac{1}{p}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_0\cap\Gamma_0}\mathcal{P}_{\mathcal{T}}\mathbf{M}-\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma_0}\mathcal{P}_{\mathcal{T}}\mathbf{M}\right)_{a,b} &= \sum_{i<j,(i,j)\in\Gamma_0}\left\langle (p^{-1}\delta_{ij}^{(k)}-1)m_{i,j}\mathcal{P}_{\mathcal{T}}\left(e_ie_j^\top+e_je_i^\top\right),\ e_ae_b^\top\right\rangle \\ &\triangleq \sum_{i<j,(i,j)\in\Gamma_0}\xi_{ij}\end{aligned}$$

where $\mathbb{E}\left[\xi_{ij}\right]=0$. We have the bounds

$$\begin{aligned}|\xi_{ij}| &\le 2p^{-1}\left\|\mathcal{P}_{\mathcal{T}}(e_ie_j^\top)\right\|_F\left\|\mathcal{P}_{\mathcal{T}}(e_ae_b^\top)\right\|_F|m_{i,j}| \\ &\le \frac{4n}{K_{\min}^2p}\left\|\mathbf{M}\right\|_\infty\end{aligned}$$

and

$$\begin{aligned}\left|\mathbb{E}\left[\sum_{i<j,(i,j)\in\Gamma_0}\xi_{ij}^2\right]\right| &= \left|\sum_{i<j,(i,j)\in\Gamma_0}\mathbb{E}\left[(p^{-1}\delta_{ij}^{(k)}-1)^2\right]m_{i,j}^2\left\langle\mathcal{P}_{\mathcal{T}}\left(e_ie_j^\top+e_je_i^\top\right),\ e_ae_b^\top\right\rangle^2\right| \\ &\le \left(p^{-1}-1\right)\left\|\mathbf{M}\right\|_\infty^2\sum_{i<j,(i,j)\in\Gamma_0}\left\langle e_ie_j^\top+e_je_i^\top,\ \mathcal{P}_{\mathcal{T}}(e_ae_b^\top)\right\rangle^2 \\ &\le 2\left(p^{-1}-1\right)\left\|\mathbf{M}\right\|_\infty^2\left\|\mathcal{P}_{\mathcal{T}}(e_ae_b^\top)\right\|_F^2 \\ &\le 2\left(p^{-1}-1\right)\frac{2n}{K_{\min}^2}\left\|\mathbf{M}\right\|_\infty^2 \\ &\le \frac{4n}{K_{\min}^2p}\left\|\mathbf{M}\right\|_\infty^2.\end{aligned}$$

When $p\ge\frac{64\beta n\log n}{3K_{\min}^2\epsilon_3^2}$ and $\epsilon_3\le 1$, we apply the standard Bernstein inequality (first inequality of (3.17)) and obtain

$$\begin{aligned}&\mathbb{P}\left[\left|\left(\frac{1}{p}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_0\cap\Gamma_0}P_{\mathcal{T}}\mathbf{M}-\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma_0}\mathcal{P}_{\mathcal{T}}\mathbf{M}\right)_{a,b}\right|\ge\epsilon_3\left\|\mathbf{M}\right\|_\infty\right] \\ \le\ & 2\exp\left(-\frac{3\epsilon_3^2\left\|\mathbf{M}\right\|_\infty^2}{8\frac{4n}{K_{\min}^2p}\left\|\mathbf{M}\right\|_\infty^2}\right)\le 2n^{-2\beta}.\end{aligned}$$

Union bound then yields

$$\mathbb{P}\left[\left\|\frac{1}{p}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_0\cap\Gamma_0}P_{\mathcal{T}}\mathbf{M}-\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma_0}\mathcal{P}_{\mathcal{T}}\mathbf{M}\right\|_\infty\ge\epsilon_3\left\|\mathbf{M}\right\|_\infty\right] \le 2n^{2-2\beta}.$$

□

The next lemma bounds the matrix infinity norm of $\mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*)$.

Lemma 25. *Under the assumption of Theorem 7 and conditioned on Ω, for some constant C_0, we have*

$$\|\mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*)\|_\infty \leq \frac{C_0 p_0}{\log n}$$

with high probability.

Proof. Under our random sign assumption and when Ω is fixed, each pair of symmetric entries of $sign(\mathbf{B}^*)$ in Ω equals ± 1 with probability $\frac{p_0}{2}$ and 0 otherwise. Notice that by triangle inequality,

$$\begin{aligned}&\|\mathcal{P}_{\mathcal{T}} sign(\mathbf{B}^*)\|_\infty \leq \left\|\mathbf{U}\mathbf{U}^T sign(\mathbf{B}^*)\right\|_\infty \\ &+ \left\|sign(\mathbf{B}^*)\mathbf{U}\mathbf{U}^T\right\|_\infty + \left\|\mathbf{U}\mathbf{U}^T sign(\mathbf{B}^*)\mathbf{U}\mathbf{U}^T\right\|_\infty,\end{aligned}$$

so it suffices to bound these three terms. Let $\left(s^{(i)}\right)^T$ be the ith row of $\mathbf{U}\mathbf{U}^T$. From the structure of $\mathbf{U}$, we know

$$\left|s_j^{(i)}\right| = \left|r_i^T r_j\right| \leq \frac{n}{K_{\min}^2}, \quad \forall i, j.$$

and

$$\sum_{j=1}^n \left(s_j^{(i)}\right)^2 = \|r_i\|_2^2 \leq \frac{n}{K_{\min}^2}, \quad \forall i.$$

Now we bound $\left\|\mathbf{U}\mathbf{U}^T sign(\mathbf{B}^*)\right\|_\infty$. For simplicity, we focus on the $(1,1)$ entry of $\mathbf{U}\mathbf{U}^T sign(\mathbf{B}^*)$ and denote this random variable as X. Observe that $X = \sum_{i:(i,1)\in\Omega} s_i^{(1)} \left(sign(\mathbf{B}^*)\right)_{i,1}$, and

$$\begin{aligned}\mathbb{E}\left[s_i^{(1)}\left(sign(\mathbf{B}^*)\right)_{i,1}\right] &= 0 \\ \left|s_i^{(1)}\left(sign(\mathbf{B}^*)\right)_{i,1}\right| &= \left|s_i^{(1)}\right| \leq \frac{n\log n}{K_{\min}^2}, \quad a.s. \\ \mathrm{Var}\left(X\right) &= \sum_{i:(i,1)\in\Omega} (s_i^{(1)})^2 p_0 \leq \frac{n}{K_{\min}^2} p_0.\end{aligned}$$

Standard Bernstein inequality (second inequality of (3.17)) thus gives

$$\mathbb{P}\left[|X| > t\right] \;\leq\; 2\exp\left(-\frac{3t}{8\frac{n\log n}{K_{\min}^2}}\right)$$

provided $t \geq \left(\frac{n}{K_{\min}^2}p_0\right) / \left(\frac{n\log n}{K_{\min}^2}\right) = \frac{p_0}{\log n}$. Choosing $t = \frac{C_0 p_0}{3\log n}$ with a suitable constant C_0 and applying the union bound, we obtain $\left\|\mathbf{U}\mathbf{U}^T sign(\mathbf{B}^*)\right\|_\infty \leq \frac{C_0 p_0}{3\log n}$ w.h.p. provided $p_0 \geq C\frac{n}{K_{\min}^2}\log^3 n$ for some C sufficiently large, which is satisfied under the assumption of Theorem 7. Similarly, we have $\left\|sign(\mathbf{B}^*)\mathbf{U}\mathbf{U}^T\right\|_\infty \leq \frac{C_0 p_0}{3\log n}$ w.h.p. under the same condition. Finally, observe that

$$\begin{aligned}\left(\mathbf{U}\mathbf{U}^T sign(\mathbf{B}^*)\mathbf{U}\mathbf{U}^T\right)_{1,1} &= \sum_{i,j:(i,j)\in\Omega} s_i^{(1)} s_j^{(1)} \left(sign(\mathbf{B}^*)\right)_{i,j} \\ &= 2\sum_{i,j:(i,j)\in\Omega} s_i^{(1)} s_j^{(1)} \left(sign(\mathbf{B}^*)\right)_{i,j}.\end{aligned}$$

Then a similar application of Bernstein inequality and the union bound gives $\left\|\mathbf{U}\mathbf{U}^T sign(\mathbf{B}^*)\mathbf{U}\mathbf{U}^T\right\|_\infty \leq \frac{C_0 p_0}{3\log n}$ w.h.p. provided $p_0 \geq C'\frac{n^2}{K_{\min}^4}\log^3 n$ for some C' sufficiently large, which is also satisfied under the assumption of Theorem 7. □

Chapter 4

Graph Clustering using Max-norm Optimization

We suggest using the max-norm as a convex surrogate constraint for clustering. We show how this yields a better exact cluster recovery guarantee than previously suggested nuclear-norm relaxation, and study the effectiveness of our method, and other related convex relaxations, compared to other clustering approaches.

4.1 Introduction

Clustering as the problem of partitioning data into clusters with strong similarity inside the clusters and strong dissimilarity across different clusters is one of the main problems in machine learning. In this chapter, we consider the problem of cut-based, or *correlation*, clustering [11], where, given a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ on n nodes with normalized symmetric affinity matrix A (for all $u, v \in \mathcal{V}$: $0 \leq A_{uv} \leq 1$ and $A_{uu} = 1$), we want to partition $\mathcal{V}$ into clusters $\mathcal{C} = \{C_1, \ldots, C_k\}$ so as to minimize the the total *disagreement*

$$D(\mathcal{C}) = \sum_{i=1}^{k} \sum_{u,v \in C_i} (1 - A_{uv}) + \sum_{i \neq j = 1}^{k} \sum_{u \in C_i, v \in C_j} A_{uv}.$$

The first term, captures the *internal* disagreement inside clusters, and the second term captures the *external* agreement between nodes in different clusters. In an ideal cluster, the affinities between all members of the same cluster are 1 and the affinities between members of two different clusters are zero and hence the objective is zero. This objective does not require the number of clusters to be known ahead of time—we may decide to use any number of clusters, and this is accounted for in the objective. Unfortunately, finding a clustering minimizing the disagreement $D(\mathcal{C})$ is NP-Hard [11].

We formulate this problem as an optimization of a convex disagreement objective over a non-convex set of *valid clustering* matrices (Section 3.3.2.1) and then consider convex relaxations of this constraint. Recently, [67] suggested a trace-norm (aka nuclear-norm) relaxation, casting the problem as minimizing an ℓ_1 loss and a trace-norm penalty, and providing conditions under which the true underlying clustering is recovered. Instead of trace-norm, we propose using the max-norm (aka $\gamma_2 : \ell_1 \to \ell_\infty$ norm) [126], which is a tighter convex relaxation than the trace-norm. Accordingly, we establish an exact recovery guarantee for our max-norm based formulation that is strictly better then the trace-norm based guarantee. We show that if the affinity matrix is a corruption of an "ideal" clustering matrix, with a certain bound on the corruption, then the optimal solution of the max-norm bounded optimization problem is exactly the ideal clustering (Section 4.3.1). We also discuss even tighter convex relaxations related to the max-norm, and suggest augmenting the convex relaxation with a single-linkage post-processing step in case of non-exact recovery, showing the empirical advantages of these approaches (Section 4.5).

The approach we suggests relies on optimizing an ℓ_1 objective subject to a max-norm constraint. A similar optimization problem with a trace-norm constraint (or trace-norm regularization) has recently been the subject of some interest in the context of "robust PCA" [24, 147] and recovering the structure of graphical models with latent variables [28]. As with the trace-norm regularized variant, the ℓ_1 + max-norm problem can be formulated as an SDP and solved using standard solvers, but this is only applicable to fairly small scale problems. In Section 4.4, we discuss various optimization approaches to this problems, including approaches which preserve the sparsity of the solution.

4.1.1 Relationship to the Goemans Willimason SDP Relaxation

Our convex relaxation approach is related to the classic SDP relaxations of max-cut [56] and more generally the cut-norm [4]. In fact, if we are interested in a partition to exactly two clusters, the correlation clustering problem is essentially a max-cut problem, though with both positive and negative weights (i.e. a symmetric cut-norm problem), and our relaxation is essentially the

classic SDP relaxation of these problems. Our approach and results differ in several ways.

First, we deal with problems with multiple clusters, and even when the number of clusters *is not* pre-determined. If the number of clusters k *is* pre-determined, the correlation clustering problem can be written as an integer quadratic program, with a k variables per node, and can be relaxed to an SDP. But this SDP will be very different from ours, and will involve a matrix of size $nk \times nk$, unlike our relaxation where the matrix is of size $n \times n$ regardless of the number of clusters. Consequently, the rounding techniques based on (random) projections typically employed for classic SDP relaxations do not seem relevant here. Instead, we employ a single-linkage post-processing as a form of "rounding" imperfect solutions.

Second, the type of guarantees we provide are very different from those in the Theory of Computation literature. Most of the SDP relaxation work we are aware of (including the classical work cited above) focuses on worst case constant factor approximation guarantees. On one hand, this means the guarantee needs to hold even on "crazy" inputs where there is really no reasonable clustering anyway, and second, and on the other hand it is not clear how approximating the objective to within a constant factor translates to recovering an underlying clustering. Instead, we prove that when the affinity matrix is close enough to following some underlying "true" clustering, the true clustering will be recovered *exactly*. This type of guarantee is more in the spirit of compressed sensing, which where *exact* recovery of a support set is guaranteed subject to conditions on the input [67].

4.1.2 Other Clustering Approaches

There are several classes of clustering algorithms with different objectives. In hierarchical clustering algorithms such as UPGMA [124], SLINK [123] and CLINK [41] the goal is to generate a sequence of clusterings by produce a sequence of clustering by merging/splitting two clusters at each step of the sequence according to a *local* disagreement objective as opposed to our global $D(\mathcal{C})$. Because of this locality, these methods are known to be very sensetive to outliers.

Cut-based clustering algorithms such as k-means/medians [65, 127], ratio association [121], ratio cut [27] and normalized cut [150] try to optimize an objective function globally. The main issue with these objectives is that they are typically NP-Hard and need to know the number of clusters ahead of time, since these objectives are monotone in the number of clusters.

In contrast, spectral clustering algorithms[141] try to find the first k principal component of the affinity matrix or a transformed version of that [96]. These methods require the number of clusters in advance and has been shown to be tractable (convex) relaxations to NP-Hard cut-based algorithms [43]. These methods are again very sensitive to outliers as they might change the principal components dramatically.

4.2 Problem Setup

Our approach is based on representing a clustering $\mathcal{C}$ through its incidence matrix $K(\mathcal{C}) \in \mathbb{R}^{n \times n}$ where $K_{uv} = 1$ iff u and v belong to the same cluster in $\mathcal{C}$ (i.e. $u, v \in C_i$ for some i), and $K_{uv} = 0$ otherwise (i.e. if u and v belong to different clusters). The matrix $K(\mathcal{C})$ is thus a permuted block-diagonal matrix, and can also be thought of as the edge incidence matrix of a graph with cliques corresponding to clusters in $\mathcal{C}$. We will say that a matrix K is a **valid clustering matrix**, or sometimes simply **valid**, if it can be written as $K = K(\mathcal{C})$ for some clustering $\mathcal{C}$ (i.e. if it is a permuted block diagonal matrix, with 1s in the diagonal blocks).

The disagreement can then be written as either:

$$D(\mathcal{C}) = \|A - K(\mathcal{C})\|_1 = \sum_{u,v} |A_{uv} - K(\mathcal{C})_{uv}| \tag{4.1}$$

or as:

$$D(\mathcal{C}) = \sum_{u,v} K(\mathcal{C})_{uv}(1 - 2A_{uv}) + \sum_{uv} A_{uv} \ , \tag{4.2}$$

where the term $\sum_{uv} A_{uv}$ does not depend on the clustering $\mathcal{C}$ and can thus be dropped.

We now phrase the correlation clustering problem as matrix problem,

where we would like to solve

$$\min_{K} D(K) \text{ s.t. } K \text{ is a valid clustering matrix.} \tag{4.3}$$

The problem is that even though the objectives (4.1) and (4.2) are convex, the constraint that K is valid is certainly not constraint. Our approach to correlation clustering will thus be to relax this non-convex constraint (the validity of K) to a convex constraint.

We note that although both the absolute error objective (4.1) and the linear objective (4.2) agree on valid clustering matrices (or more generally, on binary matrices K), they can differ when K is fractional, and especially when A is also fractional. The choice of objective can thus be important when relaxing the validity constraint to a convex constraint. More specifically, as long as A is binary (i.e. $A_{uv} \in \{0, 1\}$), and $0 \leq K_{uv} \leq 1$, even if K is fractional, the two objectives agree. Non-negativity of K_{uv} is ensured in some, but not all, of the convex relaxations we study. When non-negativity is not ensured, the absolute error objective (4.1) would tend to avoid negative values, but the linear objective might certainly prefer them. More importantly, once the affinities A_{uv} are also fractional, the two objectives differ even for $0 \leq K_{uv} \leq 1$. While the linear objective would tend to not care much about entries with affinities close to $1/2$, the absolute error objective would tend to encourage fractional values in thees cases.

The linear objective also has some optimization advantages over the absolute function as well. From a numerical optimization point of view, dealing with the linear objective function is easier since we do not need to compute the sub-gradients of the ℓ_1-norm.

4.3 Max-Norm Relaxation

As discussed in the previous Section, we are interested in optimizing over the non-convex set of valid clustering matrices. The approach we discuss here is to relaxing this set to the set of matrices with bounded *max-norm* [126]. The max-norm of a matrix K is defined as

$$\|K\|_{\max} = \min_{K=RL^T} \|R\|_{\infty,2}\|L\|_{\infty,2}$$

where, $\|\cdot\|_{\infty,2}$ is the maximum of the ℓ_2 norm of the rows, and the minimization is over factorization of any internal dimensionality. It is not hard to see that if K is a valid clustering matrix, with $K = K(\mathcal{C})$, then $\|K\|_{\max} = 1$. This is achieved, e.g., by a factorization with $R = L$, and where each row R_u of R is a (unit norm) indicator vector with $R_{ui} = 1$ for $u \in C_i$ and zero elsewhere.

Relaxing the validity constraint to a max-norm constraint, and using the absolute error objective, we obtain the following convex relaxation of the correlation clustering problem:

$$\widehat{K} \;=\; \arg\min_{K} \|A - K\|_1 \quad \text{s.t.} \quad \|K\|_{\max} \leq 1. \tag{4.4}$$

Alternatively, we could have used the linear objective (4.2) instead. In any case, after finding $\widehat{K}$, it is easy to check whether it is valid, and if so recover the clustering from its block structure. If $\hat{K}$ is valid, we are assured the corresponding clustering is a globally optimal solution of the correlation clustering problem.

4.3.1 Theoretical Guarantee

Assuming there exists an underlying true clustering, we provide a worst-case (deterministic) guarantee for exact recovery of that clustering in the presence of noise when the affinity matrix A is a binary $0-1$ matrix using absolute objective. The flavor of our result is similar to [67] for trace-norm, except that we show the max-norm constraint problem recovers the underlying clustering with larger noise comparing to trace-norm constraint. This matches our intuition that max-norm is a tighter relaxation than trace-norm for valid clustering matrices.

To present our theoretical result, we start by introducing an important quantity that our main result is based upon. Suppose $\mathcal{C}^* = \{C_1^*, \ldots, C_k^*\}$ is the underlying true clustering. For a node u and a cluster C_i^*, let $d_{u,C_i^*} = \frac{\sum_{v \in C_i^*} A_{u,v}}{|C_i^*|}$ if $u \notin C_i^*$ and $d_{u,C_i^*} = 1 - \frac{\sum_{v \in C_i^*} A_{u,v}}{|C_i^*|}$ otherwise and

$$D_{\max}(A, K) \equiv D_{\max}(A, K(\mathcal{C}^*)) = \max_{u,i} \; d_{u,C_i^*}$$

be the maximum of the disagreement ratios on the adjacency matrix. This definition is inspired by [67] but is slightly different. Notice that the larger

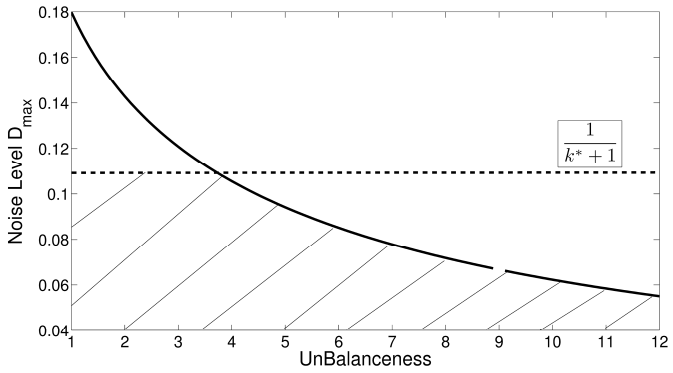


Figure 4.1: Theorem 8 guarantee region of the noise level $D_{\max}$ vs the unbalanceness parameter $\frac{1}{k^*}\sum_i \left(\frac{|C_i^*|}{|C_{\min}|}\right)^2$.

$D_{\max}(A, K)$ is, the more noisy (comparing to ideal clusters) the graph is; and hence, the harder the clustering becomes. In particular for ideal clusters (fully connected inside and fully disconnected outside clusters), we have $D_{\max}(A, K) = 0$.

We would like to ensure that when $D_{\max}(A, K)$ is small enough, our method can recover K. The following lemma helps us understand the information theoretic limit of $D_{\max}(A, K)$, i.e. what value of D_{max} is certainly *not* enough to ensure recovery, even information theoretically:

Lemma 26. *For any clustering* $\mathcal{C} = \{C_1, \ldots, C_k\}$ *and for all* $\gamma > \frac{2}{5+r}$ *with* $r = \frac{n^2}{\sum_i |C_i|^2}$, *there exists an affinity matrix* A *such that* $D_{\max}(A, K(\mathcal{C})) = \gamma$ *and the combinatorial program* (4.3) *does not output* $\mathcal{C}$.

Note that the minimum of $\frac{2}{5+r}$ is attained when all clusters have equal sizes. If we have k^* clusters of size $\frac{n}{k^*}$, then $r = k^*$ and the bound in Lemma 26 asserts that if $D_{\max}(A, K) > \frac{2}{k^*+5}$, then there are examples for which the original clustering cannot be recovered by the combinatorial program (4.3). This implies that $D_{\max}(A, K)$ cannot be scaled better than $\Theta(\frac{1}{k^*})$ in general even without convex relaxation.

Suppose there exist a true underlying clustering $\mathcal{C}^*$ with k^* clusters. Let C_{min} be the smallest size underlying true cluster and we are given an affinity

matrix A with $D_{\max} = D_{\max}(A, K(\mathcal{C}^*))$. Introducing lagrange multiplier μ, we consider the optimization problem

$$\widehat{K}_\mu \;=\; \arg\min_K \; \frac{1-\mu}{n^2}\,\|A-K\|_1 \;+\; \mu\,\|K\|_{\max}. \tag{4.5}$$

The following theorem characterizes the noise regime under which the simple max-norm relaxation (4.5) recovers $\mathcal{C}^*$.

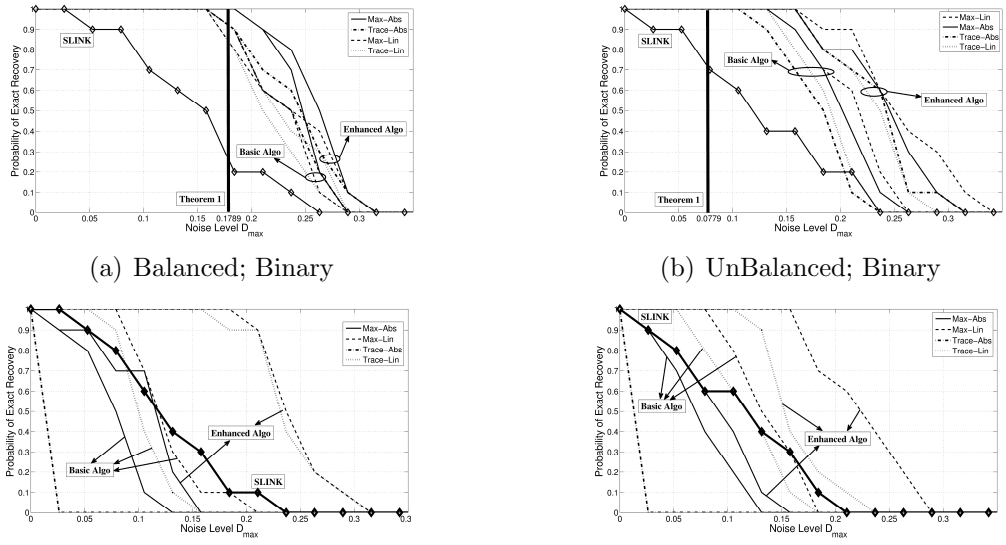
Theorem 8. *For binary $0-1$ matrix A, if $D_{\max} < \frac{1}{k^*+1}$ is small enough to satisfy $\frac{1}{k^*}\sum_i \left(\frac{|C_i^*|}{|C_{\min}|}\right)^2 \leq \frac{(1-3D_{\max})^2}{(1+D_{\max})D_{\max}}$ then, for any μ_0 satisfying $\frac{(1+D_{\max})}{(1-3D_{\max})|C_{\min}|^2} < \frac{(1-\mu_0)k^*}{\mu_0 n^2} < \frac{(1-3D_{\max})k^*}{D_{\max}\sum_i |C_i^*|^2}$, the matrix $\widehat{K}_{\mu_0}$ (the solution to (4.5)) is unique and equal to the matrix $K^* = K(\mathcal{C}^*)$ (the solution to (4.3)).*

Remark 1: Consider the parameter $\frac{1}{k^*}\sum_i \left(\frac{|C_i^*|}{|C_{\min}|}\right)^2$ in the theorem. Notice that for a balanced underlying clustering (k^* clusters of size n/k^*), this parameter is 1 and as the underlying clustering gets more and more unbalanced, this parameter increases. That motivates to call it *unbalanceness* of the clustering. It is clear that as unbalanceness parameter increases, the region of $D_{\max}$ for which our theorem guarantees the clustering recovery shrinks. We plot the admissible region of $D_{\max}$ due to unabalanceness in Fig 4.1.

Remark 2: According to the Lemma 26, the bound on $D_{\max}$ is order-wise tight and can be only improved by a constant in general.

4.3.2 Comparison to Trace-Norm Constrained Clustering

Since the max-norm constraint is strictly a tighter relaxation to the trace-norm constraint, we expect the max-norm algorithm to perform better. Our theorem shows improvement over the guarantees provided for trace-norm clustering. Comparing to the result of [67] on trace-norm ($D_{\max} \leq \frac{|C_{\min}|}{4n}$), the max-norm tolerates more noise. To see this, consider a balanced clustering, then trace-norm requires $D_{\max} \leq \frac{1}{4k^*}$ and max-norm requires $D_{\max} \leq \min(\frac{1}{k^*+1}, 0.1789)$ which is larger than $\frac{1}{4k^*}$ for all k^*. The difference gets more clear for unbalanced clustering. Suppose we have one small cluster of constant size $|C_{\min}|$ and other clusters are approximately of size $\frac{n}{k^*}$. As (n, k^*) scales, trace-norm guarantee requires that $D_{\max} = o(\frac{1}{n})$ which is inverse proportional to the size of the smallest cluster, whereas, max-norm guarantee



(a) Balanced; Binary

(b) UnBalanced; Binary

(c) Balanced; Fractional

(d) UnBalanced; Fractional

Figure 4.2: Probability of exact clustering recovery for max-norm and trace-norm constrained algorithms under absolute $\|A-K\|_1$ and linear $\sum_{i,j} K_{ij}(1-2A_{ij})$ objectives. There are 4 clusters of size 25 for the balanced case and three clusters of size 30 + one cluster of size 10 for the unbalanced case. We consider two cases for each graph; where the affinity matrix is binary and when it is not. We both show the results for simple max-norm relaxation (basic algorithm) and tighter relaxations presented in Section 4.5 (enhanced algorithm). The result shows that max-norm constrained optimization recovers the exact clustering matrix under higher noise regimes better than trace-norm and single-linkage algorithm. Also, the linear objective seems to be performing better than the absolute objective for the clustering problem in most cases.

requires $D_{\max} = o(\frac{k^*}{n})$ which is inverse proportional to the size of the largest cluster. This is a huge theoretical advantage in our theorem.

Further, we compare our algorithm with trace-norm algorithm [67] and SLINK on a probabilistic setup. Start from two different ideal clusters on 100 nodes: a) *Balanced* clusters: four ideal clusters of size 25, b) *Unbalanced* clusters: three ideal clusters of size 30 and one ideal cluster of size 10. Then, gradually increase $D_{\max}$ on both graphs and run all algorithms and report the probability of success in exact recovery of the underlying clusters. Although our theoretical guarantee is for binary affinity matrices, here, we run the same experiment for fractional affinity matrix. We run all experiments for both absolute and linear objectives. Fig. 4.3.1 shows that in all cases max-norm outperforms the trace-norm and the improvement is more significant for unbalanced clustering with fractional affinity matrix. Moreover, this experiments reveal that the absolute objective has slight advantage if the affinity matrix is binary and clusters are balanced; otherwise, the linear objective is better.

4.4 Max-norm + ℓ_1-norm Optimization

In this Section we consider optimization problems of the form (4.4). This problem recovers a sparse and low-rank matrix from their sum, considering max-norm as a proxy to rank. In Section 4.4.1, we discuss how (4.4) can be formulated as an SDP, allowing us to easily solve it using standard SDP solvers, as long as the problem size is relatively small. We then propose three other methods to numerically solve the optimization problem (4.4).

4.4.1 Semi-Definite Programming Method

Following [126], we introduce dummy variables $L, R \in \mathbb{R}^{n\times n}$ and reformulate (4.4) as the following SDP problem

$$\begin{aligned} \widehat{K} = \arg \min_{K,L,R} \ & \|A - K\|_1 \\ \text{s.t.} \quad & \begin{bmatrix} L & K \\ K^T & R \end{bmatrix} \succeq 0 \quad \text{and} \quad L_{ii}, R_{ii} \leq 1 \end{aligned}$$

These constraints are equivalent to the condition $\|K\|_{\max} \leq 1$. This SDP can be solved using generic SDP solvers, though is very slow and is not scalable to large problems.

4.4.2 Factorization Method

Motivated by [82], we introduce dummy variables $L, R \in \mathbb{R}^{n \times n}$ and let $K = LR^T$. With this change of variable, we can reformulate (4.4) as

$$\begin{aligned} \widehat{K} = \widehat{L}\widehat{R}^T = \arg\min_{L,R} \ & \|A - LR^T\|_1 \\ \text{s.t.} \quad & \|L\|_{\infty,2}, \|R\|_{\infty,2} \leq 1. \end{aligned}$$

This problem is not convex, but it is guaranteed to have no local minima for large enough size of the problem [20]. Furthermore, if we now the optimal solution $\hat{K}$ has rank at most r, we can take L, R to be $\mathbb{R}^{n \times (r+1)}$. In practice, we truncate to some reasonably high rank r even without a known gurantee on the rank of the optimal solution. To solve this problem iteratively, [82] suggest the following update

$$\begin{bmatrix} L \\ R \end{bmatrix}_{k+1} = \mathcal{P}_{\max}\left(\begin{bmatrix} L \\ R \end{bmatrix}_k + \frac{\tau}{\sqrt{k}} \begin{bmatrix} \mathbf{Sign}(A - LR^T)\ R \\ \mathbf{Sign}(A - LR^T)^T\ L \end{bmatrix}_k \right).$$

The projection $\mathcal{P}_{\max}(\cdot)$ operates on rows of L and R; if ℓ_2-norm of a row is less than one, it remains unchanged, otherwise it will be rescaled so that the ℓ_2-norm becomes one.

A possible problem with the above formulation is the lack of "sparsity" in the following sense: The ℓ_1 objective is likely to yield and optimal solution K^* with many non-zeros in $A - K^*$, i.e. where K^* is *exactly* equal to A on some of the entries. However, gradient steps on the factorization are not likely to end up in exactly sparse solutions, and we are not likely to see any such sparsity in solutions obtained by the above method.

4.4.3 Loss Function Method

There are gradient methods such as truncated gradient [79] that produce sparse solution, however, these methods cannot be applied to this problem. We introduce a surrogate optimization problem to (4.4) by adding a loss

function. For some large $\lambda \in \mathbb{R}$, solve

$$\widehat{K} = A - \widehat{Z} = \arg \min_{Z,L,R} \|Z\|_1 + \lambda \|A - Z - LR^T\|_2^2$$
$$\text{s.t.} \quad \|L\|_{\infty,2}, \|R\|_{\infty,2} \leq 1.$$

Here, the matrix Z is sparse and includes the disagreements. For sufficiently large values of λ, the loss function ensures that the matrix $A - Z$ is close to the matrix LR^T that is a bounded max-norm matrix. To solve this problem iteratively, we use the following update

$$Z_{k+1} = \mathcal{P}_{\ell_1}\left(Z_k + \frac{\tau\lambda}{\sqrt{k}}(A - Z - LR^T)_k\right)$$
$$\begin{bmatrix} L \\ R \end{bmatrix}_{k+1} = \mathcal{P}_{\max}\left(\begin{bmatrix} L \\ R \end{bmatrix}_k + \frac{\tau\lambda}{\sqrt{k}} \begin{bmatrix} (A - Z - LR^T)\ R \\ (A - Z - LR^T)^T\ L \end{bmatrix}_k\right).$$

Here, $\mathcal{P}_{\ell_1}(\cdot)$ operates on entries; if an entry has the same sign before and after the update, it remains unchanged; otherwise, it will be set to zero. Solving directly for large values of λ might cause some problems due to the finite numerical precision. In practice, we start with some small value say $\lambda = 1$ and double the value of λ after some iterations. This way, we gradually put more and more emphasis on the loss function as we get closer to the optimal point.

4.4.4 Dual Decomposition Method

Inspired by [117], we first reformulate (4.4) by introducing a dummy variable $Z \in \mathbb{R}^{n \times n}$ as follows

$$\widehat{K} = \arg \min_{Z,K} \|A - K\|_1$$
$$\text{s.t.} \quad \|Z\|_{\max} \leq 1 \qquad \text{and} \qquad Z = K.$$

Then, introducing a Lagrange multiplier $\Lambda \in \mathbb{R}^{n \times n}$, we propose the following equivalent problem:

$$\widehat{K} = \arg \max_{\Lambda} \min_{Z,K} \|A - K\|_1 + \langle\langle \Lambda,\ K - Z \rangle\rangle$$
$$\text{s.t.} \quad \|Z\|_{\max} \leq 1.$$

Here, $\langle\langle \cdot,\ \cdot \rangle\rangle$ is the trace of the product. This problem is a saddle-point convex problem in (Z, K, Λ). To solve this, we iteratively fix Λ and optimize over (K, Z) and then, using those optimal values of (K, Z), update Λ.

For a fixed Λ, the problem can be separated into two optimization problems over K and Z as

$$\widehat{K}(\Lambda) = \arg\min_{K}\ \|A - K\|_1 + \langle\langle \Lambda,\ K \rangle\rangle$$

which can be solved using factorization method discussed above, and

$$\begin{aligned} \widehat{Z}(\lambda) = \arg\min_{Z}\ & -\langle\langle \Lambda,\ Z \rangle\rangle \\ & \text{s.t.} \quad \|Z\|_{\max} \leq 1. \end{aligned}$$

which is a soft thresholding; if $|\Lambda_{ij}| > 1$ then, $\widehat{K}(\Lambda)_{ij} = -\mathbf{Sign}(\Lambda_{ij})$; otherwise $\widehat{K}(\Lambda)_{ij} = A_{ij}$.

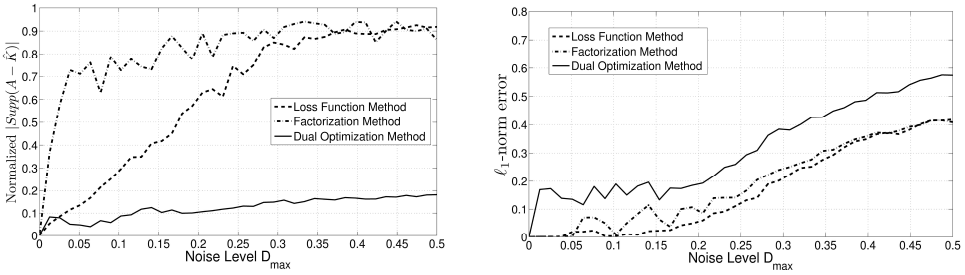
Using $\widehat{K}(\Lambda_k)$ and $\widehat{Z}(\Lambda_k)$, we update Λ as follows

$$\Lambda_{k+1} = \Lambda_k - \frac{\tau}{\sqrt{k}}(\widehat{K}(\Lambda_k) - \widehat{Z}(\Lambda_k))$$

until it converges. One criterion for the convergence of this method is to round both matrices $\widehat{K}, \widehat{Z}$ and check if they are equal. To use this criterion, we need to initialize the two matrices very differently to avoid the stopping due to the initialization.

4.4.5 Numerical Comparison

We compare the performance of these methods. For three ideal clusters of size 20 with noise level $D_{\max}$, we run all three algorithms for 2000 iterations. We consider an initial step size $\tau = 1$ for all methods, and, for the loss function method, we doubel λ every 100 iterations. For the dual method, we update Λ for 20 times and run 100 iterations of the factorization method for the max-norm sub-problem at each update. We report the sparsity of the solution $A - \widehat{K}$ as well as the ℓ_1-norm of the error $\|\widehat{K} - K^*\|_1$ for each algorithm in Fig 4.3. This result shows that there is a trade-off between sparsity and the error – the dual optimization method provides consistently a sparse solution, where, factorization and loss function methods provide small error. The sparsity of loss function method gets worse as the noise increases.



(a) $\|\text{Supp}(A - \hat{K})|$ (b) $\|K^* - \hat{K}\|_1$

Figure 4.3: Comparison of the proposed numerical optimization methods in terms of the sparsity of the solution they provide and the ℓ_1 error of the estimation.

4.5 Tighter Relaxations

In this section, we improve our basic algorithm in two ways: first, we use a tighter relaxation for valid clustering constraint and second, we add a single-linkage step after we recovered the clustering matrix. Although max-norm is a tighter relaxation comparing to trace-norm, we would like to go further and introduce tighter relaxations. Figure 4.4 summarizes different possible relaxations based on max-norm. The arrows in this figure indicated the strict subset relations among these relaxations. The tightest relaxation we suggest is $\{K = RR^T : \|R\|_{\infty,2} \leq 1, R \geq 0\}$ based on the intuition that a clustering matrix is symmetric and has a trivial factorization $R \in \mathbb{R}^{n \times k}$, where, R_{ij} is non-zero if node i belongs to cluster j. Next lemma formalizes this result.

Lemma 27. *All relaxation sets shown in Fig. 4.4 are convex and the strict subset relations hold.*

This suggests using the tightest convex relaxation, that is constraining to K such that there exists $R >= 0, \|R\|_{\infty,2} <= 1$ with $K = RR^T$ (the set of matrices K with a factorization $K = RR^T, R >= 0$ is called the set of *completely positive matrices* and is convex [16]). We optimize over this relaxation by solving the following optimization problem over R:

$$\begin{aligned} \widehat{R} \;=\; \arg\min_{R} \;& \|A - RR^T\|_1 \\ \text{s.t.} \quad & \|R\|_{\infty,2} \leq 1 \quad \& \quad R \geq 0. \end{aligned} \tag{4.6}$$

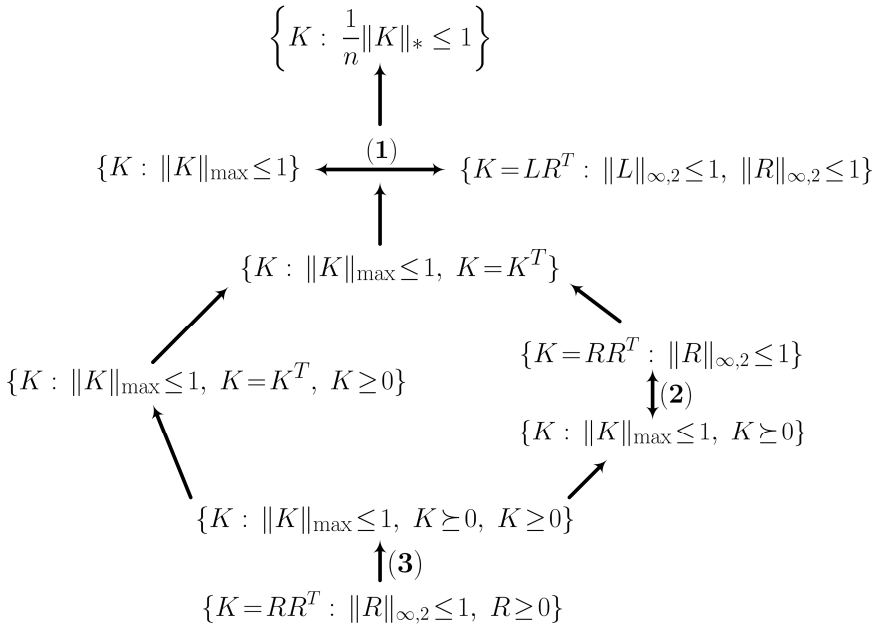


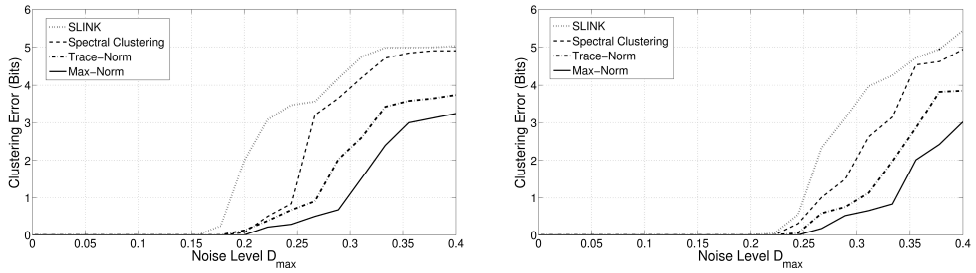
Figure 4.4: Summary of possible convex relaxations of the set of valid clustering matrices and their relations. Here, $\|\cdot\|_*$ represents the trace (nuclear) norm, $\|\cdot\|_{\infty,2}$ represents the maximum ℓ_2 norm of the rows, "$\geq$" is used for element-wise positiveness and "$\succeq$" is used for positive semi-definiteness. Each double-ended arrow represents the equivalence of two sets. Each single-ended arrow in this figure represents a *strict* sub-set relation between two sets.

and setting $\widehat{K} = \widehat{R}\widehat{R}^T$. Although the constraint on $\widehat{K}$ *is* convex, the optimization problem (4.6) is *not* convex in R.

4.5.1 Single-linkage Post Processing

The matrix $\widetilde{K}$ extracted from (4.6) might diverge from a valid clustering matrix in two ways: firstly, it might not have the structure of a valid clustering and secondly, even if it has the structure, the values might not be integer. We run SLINK on $\widetilde{K}$ as a "rounding scheme" to fix both of the above problems. SLINK gives a sequence of clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_n$. To pick the best clustering, we choose

$$\widehat{K} = \arg\min_i \|A - K(\mathcal{C}_i)\|_1. \tag{4.7}$$



(a) Balanced; Fractional (b) UnBalanced; Fractional

Figure 4.5: Comparison of our *best* proposed method which is the linear objective over tight relaxation (followed by a single-linkage algorithm) with trace-norm counterpart, single-linkage algorithm and spectral clustering. Here, we plot the entropy-based distance of the recovered clustering with the underlying true clustering.

The matrix $\widetilde{K}$ can be viewed as a refined version of the affinity matrix A and hence the second step of the algorithm can be replaced by other hierarchical clustering algorithms. The criterion of choosing the *best* clustering in the hierarchy comes naturally from the correlation clustering formulation.

4.5.2 Comparison with Other Algorithms

We compare our enhanced algorithm with the trace-norm algorithm [67] followed by SLINK and SLINK itself. In all cases we pick a clustering from SLINK hierarchy using (4.7). The setup is identical to the experiment explained in Section 4.3.2. Fig 4.3.1 summarizes the results and shows that our enhanced algorithm outperforms all competitive methods significantly.

Besides the exact recovery of the underlying clustering, we would like to investigate that as noise level $D_{\max}$ increases, how bad the output of our algorithm get. Using "variation of information" [95] as a distance measure for clusterings, we compare our algorithm with linear objective with trace-norm counterpart, SLINK and spectral clustering[141] for both balanced and unbalanced clusterings described before. For the spectral clustering method, we first find the largest $k = 4$ principal components of A and then, run SLINK on principal components. Fig 4.5 shows the result indicating that max-norm,

even when the noise level is high and no method can recover the exact clustering, outputs a clustering that is not far from the true underlying clustering in our metric.

4.6 Proof of Lemma 27

Provided equivalences (1) and (2), it is clear that $\{K = LR^T : \|L\|_{\infty,2} \leq 1, \|R\|_{\infty,2} \leq 1\}$ and $\{K = RR^T : \|R\|_{\infty,2} \leq 1\}$ are both convex sets. Since $\{K = RR^T : \|R\|_{\infty,2} \leq 1, R \geq 0\}$ is the intersection of two sets $\{K = RR^T : \|R\|_{\infty,2} \leq 1\}$ and $\mathcal{CP}\{K = RR^T : R \geq 0\}$, it suffices to show that $\mathcal{CP}$ is a convex set. The set $\mathcal{CP}$ is called the set of *completely positive matrices* and has been shown to be a closed convex cone (see Theorem 2.2 in [16] for details).

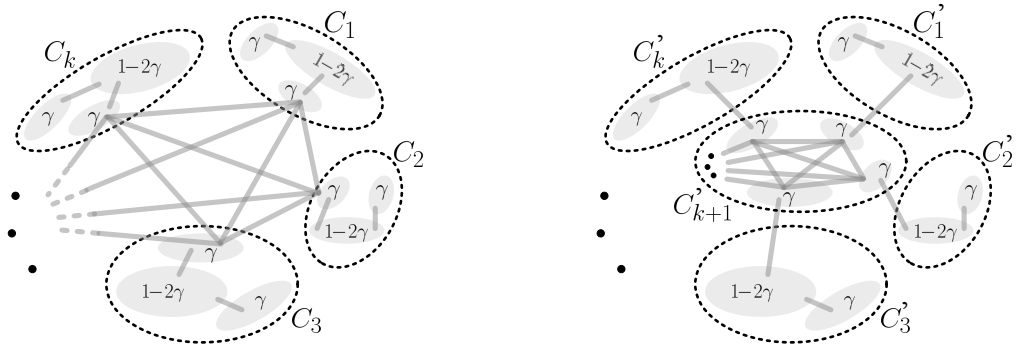
For the proof of equivalence (1) see Lemma 15 in [125]. To prove equivalence (2), it is clear that $\{K = RR^T : \|R\|_{\infty,2} \leq 1\} \subseteq \{K : \|K\|_{\max} \leq 1, K \succeq 0\}$. Now, suppose $K_0 \in \{K : \|K\|_{\max} \leq 1, K \succeq 0\}$; let $R_0 = \sqrt{K_0}$ and in contrary, assume that $\|R_0\|_{\infty,2} > 1$. This implies that at least one element on the diagonal of K_0 exceeds 1 and hence $\|K_0\|_{\max} > 1$. This is a contradiction and hence the equivalence (2) follows.

To show the relation (3), it suffices to show that the sub-set relation is strict, since the sub-set relation itself is trivial. By counter-example provided in [57], the sub-set relation is strict (i.e., there is a positive semi-definite and positive entry K_0 that does not belong to $\mathcal{CP}$).

4.7 Proof of Lemma 26

We construct an example with $D_{\max} = \frac{2}{\frac{n^2}{\sum_i |C_i|^2}+5}$ that cannot be recovered. Consider the clustering shown in Fig. 4.6(a). It is clear that for this clustering, we have $D_{\max} = \gamma$ and

$$B(\mathcal{C}_1) = \gamma^2 \sum_{i=1}^{k} |C_i|^2 + \frac{\gamma^2}{2} \sum_{i=1}^{k} |C_i|(n - |C_i|).$$



(a) Original Clustering (b) Alternative Clustering

Figure 4.6: Illustration of two alternative clusterings on the same graph with $D_{\max} = \gamma$. Each gray cloud of points is a clique. Each link between two clouds of points connects every points on one cloud to every points on the other cloud.

Now, consider the alternative clustering shown in Fig. 4.6(b). For this alternative clustering, we have

$$B(\mathcal{C}_2) = \gamma(1-2\gamma)\sum_{i=1}^{k}|C_i|^2.$$

It is clear that $B(\mathcal{C}_2) < B(\mathcal{C}_1)$ (the alternative is a better clustering) for $\gamma > \frac{2}{\frac{n^2}{\sum_i |C_i|^2}+5}$.

4.8 Proof of Theorem 8

The proof has two main steps; in the first step, we characterize a sufficient optimality condition set based on the existence of a dual variable and in the second step, we construct such dual variable. For the sake of the proof, we consider a useful equivalent definition [83] of the max norm as

$$\|K\|_{\max} = \max_{X:\|X\|_2\leq 1} \|K \circ X\|_2 \tag{4.8}$$

where, $\|\cdot\|_2$ is the spectral norm (maximum eigenvalue) of the matrix and "$\circ$" is the Hadamard element-wise product.

4.8.1 Notation

In this section, we introduce our notation and definitions used throughout the paper.

4.8.1.1 Residual Matrix Notations

In general, we do not expect the residual matrix $B^* = A - K^*$ to be sparse unless we threshold the affinity matrix (or we have adjacency matrix). However, to provide a guarantee, we need to characterize the sub-gradient of the ℓ_1-norm and hence distinguish between zeros and non-zeros of B^*. Let

$$\Omega = \{B \in \mathbb{R}^{n \times n} : B = B^T, \mathbf{Supp}(B) \subseteq \mathbf{Supp}(B^*)\}, \tag{4.9}$$

where, $\mathbf{Supp}(\cdot)$ is the index set of non-zero entries. The orthogonal projection of a matrix M to this space is defined to be a matrix of the same size with $\mathcal{P}_\Omega(M)_{ij} = M_{ij}$ if $(i,j) \in \mathbf{Supp}(B^*)$ and zero otherwise. The orthogonal complement of this space is denoted by $\Omega^\perp$ and the projection is defined as $\mathcal{P}_{\Omega^\perp}(M) = M - \mathcal{P}_\Omega(M)$.

4.8.1.2 Clustering Matrix Notations

Let $U \in \mathbb{R}^{n \times k^*}$ be constructed as

$$U = \begin{bmatrix} \frac{1}{\sqrt{|C_1|}}\mathbf{1}_{|C_1|} & & & \\ & \frac{1}{\sqrt{|C_2|}}\mathbf{1}_{|C_2|} & & \\ & & . & \\ & & & \frac{1}{\sqrt{|C_{k^*}|}}\mathbf{1}_{|C_{k^*}|} \end{bmatrix}. \tag{4.10}$$

Define $\mathcal{T} = \{UX^T + YU^T : X, Y \in \mathbb{R}^{n \times k^*}\}$ to be the space of matrices sharing either row or column space with U. The orthogonal projection to this space can be defined as

$$\mathcal{P}_\mathcal{T}(M) = UU^T M + MUU^T - UU^T MUU^T,$$

where,

$$UU^T = \begin{bmatrix} \frac{1}{|C_1|}\mathbf{1}_{|C_1|\times|C_1|} & & & \\ & \frac{1}{|C_2|}\mathbf{1}_{|C_2|\times|C_2|} & & \\ & & . & \\ & & & \frac{1}{|C_{k^*}|}\mathbf{1}_{|C_{k^*}|\times|C_{k^*}|} \end{bmatrix}.$$

Denote the orthogonal complement of the space $\mathcal{T}$ by $\mathcal{T}^\perp$ equipped with projection $\mathcal{P}_{\mathcal{T}^\perp}(M) = M - \mathcal{P}_{\mathcal{T}}(M)$. Let $\alpha = 2D_{\max}$ be the contraction between the ideal clusters and disagreements (See Lemma 30 for more details on this definition). Under the assumption of the theorem, we have $\alpha < 1$ and hence, $\mathcal{T} \cap \Omega = \{0\}$.

Using definitions in (4.11), let

$$X^* = W(Z^*) + V(UU^T),$$

where,

$$Z^* = \mathcal{P}_\Omega \begin{bmatrix} \frac{1}{|C_1|}\mathbf{1}_{|C_1|\times|C_1|} & \frac{1}{\sqrt{|C_1|\,|C_2|}}\mathbf{1}_{|C_1|\times|C_2|} & . & \frac{1}{\sqrt{|C_1|\,|C_{k^*}|}}\mathbf{1}_{|C_1|\times|C_{k^*}|} \\ \frac{1}{\sqrt{|C_2|\,|C_1|}}\mathbf{1}_{|C_2|\times|C_1|} & \frac{1}{|C_2|}\mathbf{1}_{|C_2|\times|C_2|} & . & \frac{1}{\sqrt{|C_2|\,|C_{k^*}|}}\mathbf{1}_{|C_2|\times|C_{k^*}|} \\ . & . & . & . \\ \frac{1}{\sqrt{|C_{k^*}|\,|C_1|}}\mathbf{1}_{|C_{k^*}|\times|C_1|} & \frac{1}{\sqrt{|C_{k^*}|\,|C_2|}}\mathbf{1}_{|C_{k^*}|\times|C_2|} & . & \frac{1}{|C_{k^*}|}\mathbf{1}_{|C_{k^*}|\times|C_{k^*}|} \end{bmatrix}.$$

Notice that $\mathcal{P}_{\mathcal{T}}(X^*) = UU^T$ and hence $X^* - UU^T \in \mathcal{T}^\perp$. If we show that $X^* - UU^T$ has spectral norm less than 1, then it is immediate that $X^* \in \arg\max_{X:\|X\|_2\leq 1} \|K^* \circ X\|_2$. Also, we have an eigenvalue decomposition $K^* \circ X^* = [U\ V]\Sigma[U\ V]^T$, where, U is as defined above and contains the eigenvector(s) corresponding to the maximum magnitude eigenvalue $+1$ (with k^* repetitions). To bound the spectral norm of $X^* - UU^T$, consider

$$\begin{aligned} \|X^* - UU^T\|_2 &= \|W(\mathcal{P}_\Omega(Z^* - UU^T))\|_2 \\ &\leq \frac{D_{\max}}{1-\alpha}(k^* - 1) < 1. \end{aligned}$$

The first inequality follows from Lemma 31. We make assumptions so that the last inequality holds.

We use the variational form (4.8) to characterize the sub-gradient of the max-norm at the point K^*.

Lemma 28. *For a matrix $M \in \mathbb{R}^{n\times n}$, we have $M \in \partial\|K^*\|_{\max}$ if $M = (USU^T + W) \circ X^*$, for some diagonal positive semi-definite matrix $S \in \mathbb{R}^{r\times r}$ with* $\mathbf{Trace}\,(S) = 1$ *and for some matrix $W \in \mathbb{R}^{n\times n}$ with $\mathcal{P}_{\mathcal{T}}(W) = 0$ and $\|W\|_* < 1$.*

Proof. Using the variational form (4.8) and theorem 4.4.2 in [71] on the sub-gradient of the maximum of convex functions, we have

$$\partial\|K^* \circ X^*\|_2 \subseteq \partial\|K^*\|_{\max}.$$

Thus, it suffices to show that $M \in \partial\|K^* \circ X^*\|_2$ (which is the case).

□

4.8.2 Sufficient Optimality Conditions

We provide similar optimality conditions to those provided in ℓ_1 plus trace norm minimization in the literature. The main difference here is the existence of the auxiliary variable X^* in the conditions. The following lemma characterizes a sufficient optimality condition set.

Lemma 29 (Sufficient Optimality Condition.)**.** *$K^* = \widehat{K}_\mu$ (Problem (4.3) $\equiv$ Problem (4.5)), if $\mathcal{T} \cap \Omega = \{0\}$ and there exists a dual matrix Q such that*

(a) $\mathcal{P}_\Omega(Q \circ X^*) = -\frac{1-\mu}{n^2}\mathbf{Sign}(A - K^*)$

(b) $\|\mathcal{P}_{\Omega^\perp}(Q \circ X^*)\|_\infty < \frac{1-\mu}{n^2}$

(c) $\mathcal{P}_{\mathcal{T}}(Q) = USU^T$, *for some diagonal matrix $S \succeq 0$ with Trace$(S) = \mu$.*

(d) $\|\mathcal{P}_{\mathcal{T}^\perp}(Q)\|_* < \mu$.

Proof. Notice that since X^* by construction has no zero entry (except for the very corner case where there are only two clusters both of size 2), the matrix $Q \circ X^*$ can take any value/sign on each entry by choosing the values

of Q properly. Under these conditions, $Q \circ X^* \in \partial\|A - K^*\|_1$ and also $Q \circ X^* \in \partial\|K^*\|_{\max}$ and the result follows from the standard first order optimality argument and zero duality gap of both ℓ_1 and max norms.

□

4.8.3 Dual Variable Construction

First notice that under the assumption of the theorem, we have $\alpha < 1$ and hence, by Lemma 30, we have $\mathcal{T} \cap \Omega = \{0\}$ and also $\mu = \mu_0$ is feasible. Second, we construct Q by using alternating projections. Consider the infinite sums

$$\begin{aligned} W(M) &= M - \mathcal{P}_{\mathcal{T}}(M) + \mathcal{P}_{\Omega}(\mathcal{P}_{\mathcal{T}}(M)) - \mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\Omega}(\mathcal{P}_{\mathcal{T}}(M))) + \dots \\ V(N) &= N - \mathcal{P}_{\Omega}(N) + \mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\Omega}(N)) - \mathcal{P}_{\Omega}(\mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\Omega}(N))) + \dots \end{aligned} \tag{4.11}$$

By the proof of the Lemma 30, these sums converge geometrically with parameter α (See Lemma 5 in [67] for the proof). Denoting element-wise division with "/" (and $\frac{0}{0} = 0$), let

$$Q = -\frac{1-\mu}{n^2} W(\mathbf{Sign}(A - K^*)/X^*) + \frac{\mu}{k^*} V(UU^T).$$

It is easy to check that conditions (a) and (c) in lemma 29 are both satisfied for $S = \frac{1}{k^*}\mathbf{I}$. To show condition (b), first notice that $\|\mathcal{P}_{\Omega^\perp}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}(M)\|_\infty \leq D_{\max}\|\mathcal{P}_{\Omega^\perp}(M)\|_\infty$ and hence, we have

$$\begin{aligned} \|\mathcal{P}_{\Omega^\perp}(Q \circ X^*)\|_\infty &\leq \max_i \frac{1}{(1-D_{\max})^2}\left(\frac{(1-\mu)|C_i|}{n^2} D_{\max} + \frac{\mu}{k^*}\frac{1}{|C_i|}\right)\frac{1+D_{\max}}{|C_i|} \\ &= \frac{1}{(1-D_{\max})^2}\left(\frac{(1-\mu)}{n^2}(1+D_{\max})D_{\max} + \frac{\mu}{k^*}\frac{1+D_{\max}}{|C_{\min}|^2}\right) < \frac{1-\mu}{n^2}. \end{aligned}$$

The last inequality holds for $\frac{(1-\mu)k^*}{\mu n^2} > \frac{(1+D_{\max})}{(1-3D_{\max})|C_{\min}|^2}$. For the condition (d), we have

$$\|\mathcal{P}_{\mathcal{T}^\perp}(Q)\|_* \le \frac{1}{1-\alpha}\left\|\mathcal{P}_{\mathcal{T}^\perp}\left(\frac{1-\mu}{n^2}\mathbf{Sign}(A-K^*)/X^* + \frac{\mu}{k^*}\mathcal{P}_\Omega(UU^T)\right)\right\|_*$$

$$\le \frac{D_{\max}}{1-\alpha}\frac{1-\mu}{n}\left\|\begin{bmatrix} \frac{|C_1|}{n}\mathbf{1}_{|C_1|\times|C_1|} & \frac{\sqrt{|C_1|\,|C_2|}}{n}\mathbf{1}_{|C_1|\times|C_2|} & \cdot & \frac{\sqrt{|C_1|\,|C_{k^*}|}}{n}\mathbf{1}_{|C_1|\times|C_{k^*}|} \\ \frac{\sqrt{|C_2|\,|C_1|}}{n}\mathbf{1}_{|C_2|\times|C_1|} & \frac{|C_2|}{n}\mathbf{1}_{|C_2|\times|C_2|} & \cdot & \frac{\sqrt{|C_2|\,|C_{k^*}|}}{n}\mathbf{1}_{|C_2|\times|C_{k^*}|} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\sqrt{|C_{k^*}|\,|C_1|}}{n}\mathbf{1}_{|C_{k^*}|\times|C_1|} & \frac{\sqrt{|C_{k^*}|\,|C_2|}}{n}\mathbf{1}_{|C_{k^*}|\times|C_2|} & \cdot & \frac{|C_{k^*}|}{n}\mathbf{1}_{|C_{k^*}|\times|C_{k^*}|} \end{bmatrix}\right\|_*$$

$$+ \frac{D_{\max}}{1-\alpha}\left\|\begin{bmatrix} \frac{\mu}{k^*}\frac{1}{|C_1|}\mathbf{1}_{|C_1|\times|C_1|} & \mathbf{0} & \cdot & \mathbf{0} \\ \mathbf{0} & \frac{\mu}{k^*}\frac{1}{|C_2|}\mathbf{1}_{|C_2|\times|C_2|} & \cdot & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot \\ \mathbf{0} & \mathbf{0} & \cdot & \frac{\mu}{k^*}\frac{1}{|C_{k^*}|}\mathbf{1}_{|C_{k^*}|\times|C_{k^*}|} \end{bmatrix}\right\|_*$$

$$= \frac{D_{\max}}{1-\alpha}\left(\frac{1-\mu}{n}\frac{\sum_i |C_i|^2}{n} + \mu\right) < \mu.$$

The last inequality holds for $\frac{(1-\mu)k^*}{\mu n^2} < \frac{(1-\alpha-D_{\max})k^*}{D_{\max}\sum_i |C_i|^2}$ as assumed.

Lemma 30. *If $\alpha < 1$ then $\mathcal{T} \cap \Omega = \{0\}$.*

Proof. We show that the projection $\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega(\cdot)$ has a norm α strictly less than one. Then, if there exists a non-zero matrix $M \in \mathcal{T} \cap \Omega$, then $\|M\|_\infty = \|\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega(M)\|_\infty \le \alpha\|M\|_\infty < \|M\|_\infty$ is a trivial contradiction. Let $M \in \Omega$ and consider

$$\|\mathcal{P}_\mathcal{T}(M)\|_\infty = \max_{i,j}\left\|\frac{1}{|C_i|}\mathbf{1}_{|C_i|\times|C_i|}M_{C_i,C_j} + \frac{1}{|C_j|}M_{C_i,C_j}\mathbf{1}_{|C_j|\times|C_j|} - \frac{1}{|C_i|\,|C_j|}\mathbf{1}_{|C_i|\times|C_i|}M_{C_i,C_j}\mathbf{1}_{|C_j|\times|C_j|}\right\|_\infty$$
$$\le 2D_{\max}\|M\|_\infty = \alpha\|M\|_\infty.$$

The last step is attained by optimizing over $|C_i|$ and $|C_j|$. This concludes the proof of the lemma.

□

Lemma 31. $\|W(\mathcal{P}_\Omega(Z^* - UU^T))\|_2 \le \frac{D_{\max}}{1-\alpha}(k^*-1)$.

Proof. For $M \in \Omega$, we have $\|M\|_2 \le \|M_\sigma\|_2$, where, $M_\sigma \in \mathbb{R}^{k^*\times k^*}$ with $(M_\sigma)_{i,j} = \|M_{C_i,C_j}\|_2$. By definition of $D_{\max}$, we have $\|M_{C_i,C_j}\|_2 \le D_{\max}\sqrt{|C_i|\,|C_j|}\|M_{C_i,C_j}\|_\infty$.

Thus,

$$\|\mathcal{P}_\Omega(Z^* - UU^T)\|_2 \leq D_{\max} \left\| \begin{bmatrix} 0 & 1 & \cdot & 1 \\ 1 & 0 & \cdot & 1 \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & \cdot & 0 \end{bmatrix} \right\|_2$$
$$= D_{\max}(k^* - 1).$$

The rest of the proof is straight forward as follows

$$\|W(\mathcal{P}_\Omega(Z^* - UU^T))\|_2 = \left\| \mathcal{P}_{\mathcal{T}^\perp} \mathcal{P}_\Omega \left(\sum_{i=0}^{\infty} (\mathcal{P}_{\mathcal{T}} \mathcal{P}_\Omega)^i (\mathcal{P}_\Omega(Z^* - UU^T)) \right) \right\|_2$$
$$\leq \left\| \mathcal{P}_\Omega \left(\sum_{i=0}^{\infty} (\mathcal{P}_{\mathcal{T}} \mathcal{P}_\Omega)^i (\mathcal{P}_\Omega(Z^* - UU^T)) \right) \right\|_2$$
$$= \frac{D_{\max}}{1 - \alpha}(k^* - 1).$$

This concludes the proof of the lemma.

□

Chapter 5

Learning the Dependence Graph of Time Series with Latent Factors

This chapter considers the problem of learning, from samples, the dependency structure of a system of linear stochastic differential equations, when some of the variables are *latent*. In particular, we observe the time evolution of some variables, and never observe other variables; from this, we would like to find the dependency structure between the observed variables – *separating out* the spurious interactions caused by the (marginalizing out of the) latent variables' time series. We develop a new method, based on convex optimization, to do so in the case when the number of latent variables is smaller than the number of observed ones. For the case when the dependency structure between the observed variables is sparse, we theoretically establish a high-dimensional scaling result for structure recovery. We verify our theoretical result with both synthetic and real data (from the stock market).

5.1 Introduction

Motivated by finance applications, time-series forecasting has got a lot of attention during the past three decades [33]. In the model based approaches, it is assumed that the time-series evolves according to some statistical model such as linear regression model [18], transfer function model [19], vector autoregressive model [144], etc. In each case, researchers have developed different methods to learn the parameters of the model for the purpose of forecasting. In this chapter, we focus on linear stochastic dynamical systems that are an instance of vector autoregressive models. Previous work toward estimating this model parameters include ad-hoc use of neural network [6] or support vector machine method [76], all without providing theoretical guarantees on

the performance of the algorithm.

Linear stochastic dynamical systems are classic processes that are widely used to model time series data in a huge number of fields: financial data [38], biological networks of species [81] or genes [12], chemical reactions [55, 61], control systems with noise [149], etc. An important task in several of these domains is learning the model from data [139]; doing so is often the first step in both data interpretation, and making predictions of future values or the effect of perturbations. Often one is interested in learning the *dependency structure* [72]; i.e. identifying, for each variable, which set of other variables it directly interacts with. For stock market data, for example, this can reveal which other stocks most directly affect a given stock.

We consider model structure learning in a particularly challenging yet widely prevalent setting: where (the time series of) some state variables are observed, and others are *unobserved/latent*. We are interested in learning the dependency structure between the observed variables. However, the presence of latent time series, if not properly accounted for by the model learning procedure, will result in the appearance of spurious interactions between observed variables – two observed variables that interact with the same unobserved variable may now be reported to be interacting. This happens, for example, if one uses the classic maximum-likelihood estimator [51], and persists even if we have observations over a long time horizon.

Suppose, for illustration, that we are interested in learning the dependency structure between the prices of a set of stocks via a linear stochastic model. Clearly, stock prices depend not only on each other, but are also jointly influenced by several variables that may not be part of our model, for example, currency markets, commodity prices etc.; these are latent time series. Their presence means that a naive structure learning algorithm (say max-likelihood) that takes as input only the stock prices, will report several spurious interactions; say, e.g. between all stocks that fluctuate with the price of oil.

Our work involves several significant differences from the large body of work on sparse recovery and graphical model learning. One is the fact that our samples are dependent on each other, with the degree of dependence governed by how finely the system is sampled. Another is the presence of latent variables.

Clearly there are several issues with regards to fundamental identifiability, and sample and computational complexity, that need to be defined and resolved. We do so below in the specific context of our model setting. We provide both theoretical characterization and guarantees of the problem, as well as numerical illustrations for both synthetic data and some real data extracted from stock market.

5.2 Problem Setting and Main Idea

This paper considers the problem of structure learning in linear stochastic dynamical systems, in a setting where only a subset of the time series are observed, and others are unobserved/latent. In particular, we consider a system with state vectors $x(t) \in \mathbb{R}^p$ and $u(t) \in \mathbb{R}^r$, for $t \in \mathbb{R}^+$ and dynamics described by

$$\frac{d}{dt}\begin{bmatrix} x(t) \\ u(t) \end{bmatrix} = \underbrace{\begin{bmatrix} A^* & B^* \\ C^* & D^* \end{bmatrix}}_{\mathcal{A}^*} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} + \frac{d}{dt}w(t), \tag{5.1}$$

where, $w(t) \in \mathbb{R}^{p+r}$ is an independent standard Brownian motion vector and A^*, B^*, C^*, D^* are system parameters.

Task: We observe the process $x(t)$ for some time horizon $0 \leq t \leq T$, but not the process $u(\cdot)$. We are interested in learning the matrix A^*, which captures the interactions between the observed variables.

We will also be interested in a similar objective for an analogous *discrete time* system with parameter $0 < \eta < \frac{2}{\sigma_{\max}(\mathcal{A}^*)}$:

$$\begin{bmatrix} x(n+1) \\ u(n+1) \end{bmatrix} - \begin{bmatrix} x(n) \\ u(n) \end{bmatrix} = \eta \begin{bmatrix} A^* & B^* \\ C^* & D^* \end{bmatrix} \begin{bmatrix} x(n) \\ u(n) \end{bmatrix} + w(n) \tag{5.2}$$

for all $n \in \mathbb{N}_0$. Here, $w(n)$ is a zero-mean Gaussian noise vector with covariance matrix $\eta I_{(p+r)\times(p+r)}$. The prameter η can be thought of as the sampling step; in particular notice that as $\eta \to 0$, we recover model (5.1) from model (5.2). The upper bound on η ensures the stability of the discrete time system as required by our theorem. Intuitively, $\sigma_{\max}(\mathcal{A}^*)$ corresponds to the fastest convergence

rate in the system and the upper bound on η corresponds to the Nyquist minimum sampling rate required for the reconstruction of the signal. As done in [15], our proofs will initially focus on the discrete case (5.2), and derive results for (5.1) afterwards.

(A1) Stable Overall System: We only consider stable systems. In fact, we impose an assumption slightly stronger than the stability on the overall system. For the continuous system (5.1), we require $D := -\lambda_{\max}(\frac{\mathcal{A}^* + \mathcal{A}^{*T}}{2}) > 0$. With slightly abuse of notation, for the discrete system (5.2), we require $D := \frac{1-\Sigma_{\max}^2}{\eta} > 0$, where, $\Sigma_{\max} := \sigma_{\max}(I + \eta\mathcal{A}^*)$. ■

As a consequence of this assumption, by Lyapunov theory, the continuous system (5.1) has a unique stationary measure which is a zero-mean Gaussian distribution with positive definite (otherwise, it is not unique) covariance matrix $\mathcal{Q}^* \in \mathbb{R}^{(p+r)\times(p+r)}$ given by the solution of $\mathcal{A}^*\mathcal{Q}^* + \mathcal{Q}^*\mathcal{A}^{*T} + I = 0$. Similarly, for the discrete time system (5.2), we have $\mathcal{A}^*\mathcal{Q}^* + \mathcal{Q}^*\mathcal{A}^{*T} + \eta\mathcal{A}^*\mathcal{Q}^*\mathcal{A}^{*T} + I = 0$. This matrix $\mathcal{Q}^*$ has the form $\mathcal{Q}^* = [Q^* \; R^{*T}; \; R^* \; P^*]$, where, Q^* and P^* are the steady-state covariance matrices of the observed and latent variables, respectively, and R^* is the steady-state cross-covariance between observed and latent variables. By stability, $\mathcal{C}_{\min} := \Lambda_{\min}(\mathcal{Q}^*) > 0$ and $\mathcal{D}_{\max} := \Lambda_{\max}(\mathcal{Q}^*) < \infty$.

Identifiability: Clearly, the above objective of identifying A^* is in general impossible without some additional assumptions on the model; in particular, several different choices of the overall model (including different choices of A^*) can result in the same *effective* model for the $x(\cdot)$ process. $x(\cdot)$ would then be statistically identical under both models, and correct identification would not be possible even over an infinite time horizon. Additionally, it would in general be impossible to achieve identification if the number of latent variables is comparable to or exceeds the number of observed variables. Thus, to make the problem well-defined, we need to restrict (via appropriate assumptions) the set of models of interest.

5.2.1 Main Idea

Consider the discrete-time system (5.2) in steady state and suppose, for a moment, that we ignored the fact that there may be latent time series; in

this case, we would be back in the classical setting, for which the (population version of) the likelihood is

$$\mathcal{L}(A) = \frac{1}{2\eta^2}\mathbb{E}\left[\|x(i+1) - x(i) - \eta A x(i)\|_2^2\right].$$

Lemma 32. *For $x(\cdot)$ generated by (5.2), the the optimum $\widehat{A} := \max_A \mathcal{L}(A)$ is given by*

$$\widehat{A} \;=\; A^* + B^* R^* (Q^*)^{-1}.$$

Thus, the optimal $\widehat{A}$ is a sum of the original A^* (which we want to recover) and the matrix $B^*R^*(Q^*)^{-1}$ that captures the spurious interactions obtained due to the latent time series. Notice that the matrix $B^*R^*(Q^*)^{-1}$ has the rank at most equal to number r of latent time series. We will assume that the number of latent time series is smaller than the number of observed ones – i.e. $r < p$ – and hence $B^*R^*(Q^*)^{-1}$ is a *low-rank matrix*.

5.2.2 Identifiability

Besides identifying the effect of the latent time series, we would need the true model to be such that A^* is uniquely identifiable from $B^*R^*(Q^*)^{-1}$. We choose to study models that have a *local-global structure* where *(a)* each of the observed time series $x_i(t)$ interacts with only a few other observed series, while *(b)* each of the latent series interacts with a (relatively) large number of observed series. In the stock market example, for instance, this would model the case where the latent series corresponds to macro-economic factors, like currencies or the price of oil, that affect a lot of stock prices.

In particular, let s be the maximum number of non-zero entries in any row or column of A^* ; it is the maximum number of other observed variables any given observed variable directly interacts with. Note that this means A^* is a *sparse* matrix. Let $L^* := B^*R^*(Q^*)^{-1}$ and assume it has SVD $L^* = U^*\Sigma^*V^{*T}$, and recall that its rank is r. Then, following [35], L^* is said to be *μ-incoherent* if $\mu > 0$ is the smallest real number satisfying

$$\max_{i,j}(\|U^{*T}\mathbf{e}_i\|, \|V^{*T}\mathbf{e}_j\|) \leq \sqrt{\frac{\mu r}{p}} \quad , \quad \|U^*V^{*T}\|_\infty \leq \sqrt{\frac{r\mu}{p^2}},$$

where, $\mathbf{e}_i$'s are standard basis vectors and $\|\cdot\|$ is vector 2-norm. Smaller values of μ mean the row/column spaces make larger angles with the standard bases, and hence the resulting matrix is more dense.

(A2) Identifiability: We require that the s of the sparse matrix A^* and the μ of the low-rank L^*, which has rank r, satisfy $\alpha := 3\sqrt{\frac{\mu r}{p}} < 1$. ■

5.2.3 Algorithm

Recall that our task is to recover the matrix A^* given observations of the $x(\cdot)$ process. We saw that the max-likelihood estimate (in the population case) was the sum of A^* and a low-rank matrix; we subsequently assumed that A^* is sparse. It is natural to use the max-likelihood as the loss function for the *sum* of a sparse and low-rank matrix, and separate appropriate regularizers for each of the components. Thus, for the continuous-time system observed up to time T, we propose solving

$$(\widehat{A}, \widehat{L}) = \arg\min_{A,L} \frac{1}{2T}\int_{t=0}^{T} \|(A+L)x(t)\|_2^2 \, dt - \frac{1}{T}\int_{t=0}^{T} x(t)^T (A+L)^T dx(t) + \lambda_A \|A\|_1 + \lambda_L \|L\|_*, \tag{5.3}$$

and for the discrete-time system given n samples, we propose solving

$$(\widehat{A}, \widehat{L}) = \operatorname*{argmin}_{A,L} \frac{1}{2\eta^2 n} \sum_{i=0}^{n-1} \|x(i+1) - x(i) - \eta(A+L)x(i)\|_2^2 + \lambda_A \|A\|_1 + \lambda_L \|L\|_*. \tag{5.4}$$

Here $\|\cdot\|_1$ is the ℓ_1 norm (a convex surrogate for sparsity), and $\|\cdot\|_*$ is the nuclear norm (i.e. sum of singluar values, a convex surrogate for low-rank). The optimum $\widehat{A}$ of (5.4) or (5.3) is our estimate of A^*, and our main result provides conditions under which we recover the support of A^*, as well as a bound on the error in values $\|\widehat{A} - A^*\|_\infty$ (maximum absolute value). We provide a bound on the error $\|\widehat{L} - L^*\|_2$ (spectral norm) for the low-rank part. Notice that the discrete objective function goes to the continuous one as $\eta \to 0$.

5.2.4 High-dimensional setting

Note that when A^* is a sparse matrix, the actual degrees of freedom between the observed variables is smaller than that evinced by the ambient dimension p. Indeed, we will be interested in recovering A^* with a number of samples n that is potentially much smaller than p (for small s). In the special case when we are in steady state and $L = 0$ (i.e. λ_L large) the recovery of each row of A^* is akin to a LASSO [132] problem (of sparse vector recovery from noisy linear measurements) with Q^* being the covariance of the design matrix. We thus require Q^* to satisfy incoherence conditions that are akin to those in LASSO (see e.g. [143] for the necessity of such conditions).

(A3) Incoherence: To control the effect of the *irrelevant* (not latent) variables on the set of *relevant* variables, we require

$$\theta := 1 - \max_k \|Q^*_{\mathcal{S}_k^c \mathcal{S}_k} \left(Q^*_{\mathcal{S}_k \mathcal{S}_k}\right)^{-1} \|_{\infty,1} > 0,$$

where, $\mathcal{S}_k$ is the support of the k^{th} row of A^* and $\mathcal{S}_k^c$ is the complement of that. The norm $\|\cdot\|_{\infty,1}$ is the maximum of the ℓ_1-norm of the rows. ■

5.2.5 Related Work

Doing a thorough review of the entire body of work on sparse recovery and graphical model structure learning is beyond the scope of this section. We focus instead on the two most closely related treads of work.

First, [15] consider the problem of learning dependence graph for time series, *without* any latent variables. They implement the LASSO; the main contribution is characterizing sample complexity in the presence of sample dependence. In our setting, with latent variables, their method returns several spurious edges caused by marginalization, allowing the learning of neither A^* nor L^*.

Second, [29] considers the problem of learning the structure of gaussian graphical models with latent variables; they use the idea of sparse and low-rank decomposition of matrices [23, 25, 30, 35, 155]. However, they cannot handle dependent samples. Perhaps equally important, their focus is on the recovery of the number of latent vairables – i.e. the rank of the low-rank matrix – for which they require a large number of independent samples. We show that the alternate objective of find the direct dependence graph between the observed

variables – i.e. the support of A^* – can be done with $O(\log p)$ dependent sampling.

5.3 Main Results

In this section, we present our main result for both Continuous and Discrete time systems. We start by imposing some assumptions on the regularizers and the sample complexity.

(A4) Regularizers: We need to impose some assumptions on the regularizers to be able to guarantee our result. Let

$$m = \max\left(\frac{80}{\sqrt{D}}\|B^*\|_{\infty,1}, \sqrt{\|x(0)\|_2^2 + \|u(0)\|_2^2 + (\sqrt{\eta}+1)^2}\right),$$

be the constant capturing the effect of initial condition and latent variables through matrix B^*. We impose the following assumptions on the regularizers:

(A4-1) $\lambda_A = \frac{16m(4-\theta)}{\theta\sqrt{D}}\sqrt{\frac{\log\left(\frac{4((s+2r)p+r^2)}{\delta}\right)}{n\eta}}$.

(A4-2) $\frac{\lambda_L}{\lambda_A\sqrt{p}} = \frac{1}{1-\alpha}\left(\left(\frac{3\alpha\sqrt{s}}{4} + \frac{(8-\theta)s}{\theta(4-\theta)}\right)\left(\frac{\theta\sqrt{p}}{9s\sqrt{s}}+1\right)+\frac{1}{2}\right)$.

Note: In practice, we let $\lambda_A = c\sqrt{\log\left(4((s+2r)p+r^2)/\delta\right)/n\eta}$ and $\lambda_L = d\sqrt{p}\lambda_A$, with the constants c, d chosen by cross-validation over prediction performance.

(A5) Sample Complexity: In our setting, samples are dependent; in particular, the smaller the η the more dependent two subsequent samples. Sample complexity is thus governed by the total time horizon $\eta n = T$ over which we observe the system, and not simply n; indeed finer sampling (i.e. smaller η) requires a larger number of samples. For a probability of failure δ, we require

$$T = n\eta \geq \frac{K\ s^3}{D^2\theta^2\mathcal{C}_{\min}^2}\log\left(\frac{4((s+2r)p+r^2)}{\delta}\right).$$

Here, K is a constant independent of any other system parameter; for example, $K \geq 3\times 10^6$ suffices.

The above T is required to ensure that the empirical covariance matrix is close to the steady-state Q^*, R^*. Of course the constraint $\eta < 2/\sigma_{max}(\mathcal{A}^*)$ ensures that the sampling intervals cannot be too large.

Let $\nu = \frac{\alpha\theta}{2\mathcal{D}_{\max}} + \frac{(8-\theta)\sqrt{s}}{\mathcal{C}_{\min}(4-\theta)}$. Identifying the distance between the span spaces of $\widehat{L}$ and L^* with parameter $\rho := \min\left(\frac{\alpha}{4}, \frac{\theta\alpha\lambda_A}{5\theta\alpha\lambda_A + 16\mathcal{D}_{\max}\|L^*\|_2}\right)$, the following (unified) theorem states our main result for both discrete and continuous time systems.

Theorem 9. *If assumptions (A1)-(A5) are satisfied, then with probability* $1-\delta$*, our algorithm outputs a pair* $(\widehat{A}, \widehat{L})$ *satisfying*

(a) Subset Support Recovery: $Supp(\widehat{A}) \subset Supp(A^*)$.

(b) Error Bounds:

$$\|\widehat{A} - A^*\|_\infty \leq \nu\lambda_A \quad \textit{and} \quad \|\widehat{L} - L^*\|_2 \leq \frac{\rho}{1-5\rho}\|L^*\|_2.$$

(c) Exact Signed Support Recovery: *If additionally we have that the smallest magnitude* A_{min} *of a non-zero element of* A^* *satisfies* $A_{min} > \nu\lambda_A$*, then we obtain full signed-support recovery* $Sign(\widehat{A}) = Sign(A^*)$.

Note: Note that λ_A, as defined in **(A4-1)**, depends on the sample complexity T, and goes to 0 as T becomes large. Thus it is possible to get exact signed support recovery by making T large.

Remark 1: Our result shows that, in sparse and low-rank decomposition for latent variable modeling, recovery of only the sparse component seems to be possible with much fewer samples – $O(s^3 \log p)$ – as compared to, for example, the recovery of the exact rank of the low-rank part; the latter was show to require $\Theta(p)$ samples in [29].

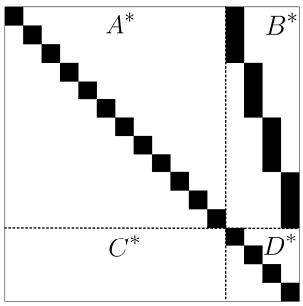
Remark 2: The above theorem shows that, even in the presence of latent variables, our algorithm requires a similar number of samples (i.e. upto universal constants) as previous work [15] required in the absence of hidden variables. Of course, this is true as long as identifiability **(A2)** holds. Note that the absence of such identifiability conditions makes even simple sparse and low-rank matrix decomposition [30] ill-posed. Note also that the quantity

ρ, which characterizes the error in the low-rank term, goes to 0 as T increases (which decreases λ_A).

Remark 3: Although our theoretical result shows a scaling proportional to s^3 for the sample complexity, the theoretical result suggests that the correct scaling factor is s^2. We suspect our result as well as [15], can be tightened.

Illustrative Example: Consider a simple idealized example that helps give intuition about the above theorem. Suppose that we are in the continuous time setting, where each latent variable j depends only on its own past, updating according to $\frac{dx_j}{dt} = -x_j(t) + \frac{dw_j}{dt}$ and for each observed variable i depends only on its own past and a *unique* latent variable $j(i)$, i.e., $\frac{dx_i}{dt} = -x_i(t) + x_{j(i)}(t) + \frac{dw_i}{dt}$. There are r latent variables, and assume that each latent variable affects exactly $\frac{p}{r}$ observed variables in this way.

In terms of the matrix $\mathcal{A}^*$, the overall (observed + latent) system has the form given by the matrix below



Here $A^* = -I_{p\times p}$, $C^* = 0$, and $D^* = -I_{r\times r}$. This matrix satisfies stability assumption **(A1)**. In the matrix B^* , each column has exactly $\frac{p}{r}$ entries that are 1, and the remaining are 0. Each row of B^* has exactly one entry that is 1, and the remaining are 0; note that the columns of B are orthogonal. We start from zero initial condition with $\eta = 0$ (continuous time system). With this, $D = 2$ and $\|B^*\|_{\infty,1} = 1$.

For this idealized setting, we can exactly evaluate all the quantities we need. In particular, it is not hard to show (done in Appendix) that the steady-state covariance matrices are $Q^* = \frac{1}{2}(I + BB^T)$ and $R^* = B^{*T}$. The resulting low-rank matrix is $L^* = \frac{r}{p+r}BB^T$, which gives $U = V = \sqrt{\frac{r}{p}}B$; the incoherence parameter $\mu = r$, and hence we need $r < \sqrt{p}/3$ by assumption

(A2). Moreover, we can show that $\theta = \frac{1}{2}$ for this example and hence the assumption **(A3)** is also satisfied.

Similarly. evaluating the other parameters in Theorem 9, we get that the observation time should be $T \geq Ks^3 \log \frac{4(1+2r)p+4r^2}{\delta}$ for structure recovery with probability greater than δ. In this case, we also have $\nu = \frac{3r}{4\sqrt{p}} + \frac{25\sqrt{s}}{7}$ and $\rho = \frac{1}{5+\frac{32\sqrt{p}}{3r\lambda_A}}$ providing the error bounds $\|A^* - \hat{A}\|_\infty \leq \left(\frac{3r}{4\sqrt{p}} + \frac{25\sqrt{s}}{7}\right)\lambda_A$ and $\|L^* - \hat{L}\|_2 \leq \frac{3r}{32\sqrt{p}}\lambda_A$.

5.4 Proof of the Theorem

In this section, we first introduce some notations and definitions and then, provide a three step proof technique to prove the main theorem for the discrete time system. The proof of the continuous time system is done via a coupling argument in the appendix.

Before we proceed to the details, we would like to make a high level technical remark on the novelties of our proof. There are two key novel ingredients in the proof enabling us to get the low sample complexity result in our theorem. The first ingredient comes from our new set of optimality conditions inspired by [23]. This optimality conditions enable us to certify an approximation of L^* while certifying the exact sign support of A^*. The second ingredient comes from the bounds on the Schur complement of the perturbation of positive semi-definite matrices [128]. This result enables us to get a bound on the Schur complement of a perturbation of a positive semi-definite matrix of size p with only $\log(p)$ samples.

Given a matrix A^*, let Ω be the subspace of matrices whose their support is a subset of the matrix A^*. The orthogonal projection of a matrix M to Ω is denoted by $\mathcal{P}_\Omega(M)$. Denote the orthogonal complement space with Ω^c with orthogonal projection $\mathcal{P}_{\Omega^c}(M)$.

For any matrix $L \in \mathbb{R}^{p\times p}$, if the SVD is $L = U\Sigma V^T$, then let $\mathcal{T}(L) := \{M | M = UX^T + YV^T \text{for some } X, Y\}$ denote the subspace spanned by all matrices that have the same column space or row space as L. The orthogonal projection of a matrix N to $\mathcal{T}$ is denoted by $\mathcal{P}_\mathcal{T}(N)$. Denote the orthogonal complement space with $\mathcal{T}^c$ with orthogonal projection $\mathcal{P}_{\mathcal{T}^c}$. We define a metric

to measure the *closeness* of two subspaces $\mathcal{T}_1$ and $\mathcal{T}_2$ as follows

$$\rho\left(\mathcal{T}_1, \mathcal{T}_2\right) = \max_{N \in \mathbb{R}^{p \times p}} \frac{\|\mathcal{P}_{\mathcal{T}_1}(N) - \mathcal{P}_{\mathcal{T}_2}(N)\|_2}{\|N\|_2}.$$

Finally, let $\mathcal{T} = \mathcal{T}(L^*)$ to shorten the notation and $L^* = U^*\Sigma^*V^*$ be a singular value decomposition.

5.4.1 Proof Technique

We outline the proof in three steps as follows:

- **STEP 1:** Constructing a candidate primal optimal solution $(\widetilde{A}, \widetilde{L})$ with the desired sparsity pattern using the restricted support optimization problem, called *oracle problem*:

$$(\widetilde{A}, \widetilde{L}) = \arg \min_{\substack{L:\rho(\mathcal{T}(L),\mathcal{T}) \leq \rho \\ A:\mathcal{P}_{\Omega^c}(A)=0}} \frac{1}{2\eta^2 n} \sum_{i=0}^{n-1} \|x(i+1) - x(i) - \eta(A+L)x(i)\|_2^2 + \lambda_A \|A\|_1 + \lambda_L \|L\|_*. \tag{5.5}$$

  This oracle is similar to the one used in [29]. It ensures that the right sparsity pattern is chosen for $\widetilde{A}$ and the tangent spaces $\widetilde{L}$ and L^* come from are *close* with parameter ρ. Note that this is a proof technique, not a method to construct the solution.

- **STEP 2:** Writing down a set of sufficient (stationary) optimality conditions for $(\widetilde{A}, \widetilde{L})$ to be the unique solution of the (unrestricted) optimization problem (5.4):

  Lemma 33. *If $\Omega \cap \mathcal{T} = \{0\}$, then $(\widetilde{A}, \widetilde{L})$, the solution to the oracle problem* (5.5)*, is the unique solution of the problem* (5.4) *if there exists a matrix $\widetilde{Z} \in \mathbb{R}^{p \times p}$ such that*

  (C1) $\mathcal{P}_\Omega(\widetilde{Z}) = \lambda_A Sign\left(\widetilde{A}\right)$.

  (C2) $\left\|\mathcal{P}_{\Omega^c}(\widetilde{Z})\right\|_\infty < \lambda_A$.

  (C3) $\left\|\mathcal{P}_{\mathcal{T}}(\widetilde{Z}) - \lambda_L U^* V^{*T}\right\|_2 \leq 4\rho\lambda_L$.

  (C4) $\left\|\mathcal{P}_{\mathcal{T}^c}(\widetilde{Z})\right\|_2 < (1-\alpha)\lambda_L$.

$$\textbf{(C5)}\ \underbrace{-\frac{1}{\eta n}\sum_{i=1}^{n}\Big(x(i+1)-x(i)-\eta(\widetilde{A}+\widetilde{L})x(i)\Big)\,x(i)^T}_{J_n}+\widetilde{Z}=0.$$

- **STEP 3:** Constructing a dual variable $\widetilde{Z}$ that satisfies the sufficient optimality conditions stated in Lemma 33. First notice that under assumption **(A2)**, we have $\Omega \cap \mathcal{T} = \{0\}$ [35]. For matrices $M \in \Omega$ and $N \in \mathcal{T}$, let

$$\begin{aligned}\mathcal{H}_M &= M - \mathcal{P}_{\mathcal{T}}(M) + \mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}(M) - \mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}(M) + \dots \\ \mathcal{G}_N &= N - \mathcal{P}_{\Omega}(N) + \mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}(N) - \mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}(N) + \dots .\end{aligned}$$

It has been shown in [35] that if $\alpha < 1$ then both infinite sums converge. Suppose we have the SVD decomposition $\widetilde{L} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$. Let

$$\widetilde{Z} = \mathcal{H}_{\lambda_A \mathrm{Sign}(\widetilde{A})} + \mathcal{G}_{\mathcal{P}_{\mathcal{T}}(\lambda_L \widetilde{U}\widetilde{V}^T)} + \Delta,$$

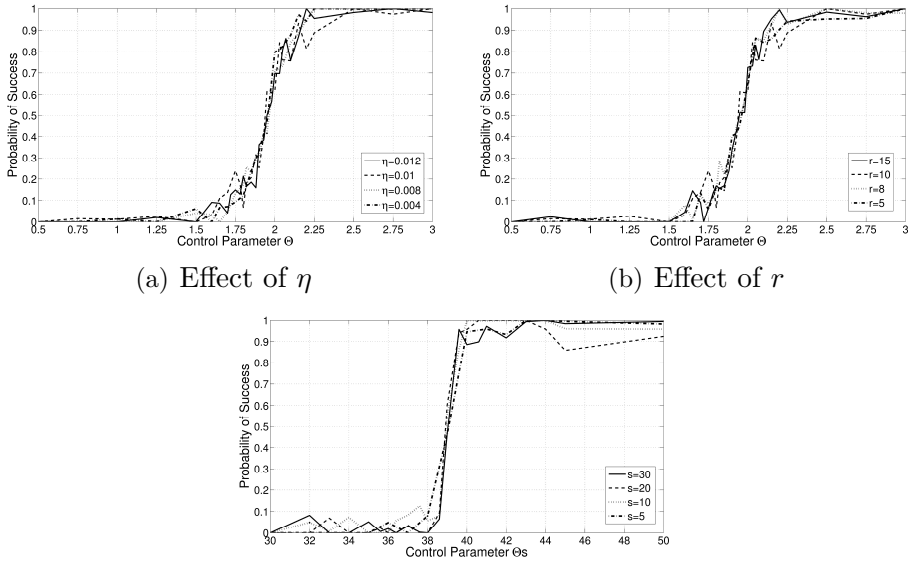
where, Δ is a matrix such that **(C5)** is satisfied. As a result of our construction, we have $\mathcal{P}_{\Omega}(\widetilde{Z} - \Delta) = \lambda_A \mathrm{Sign}(\widetilde{A})$ and by optimality of $(\widetilde{A}, \widetilde{L})$, we have $\mathcal{P}_{\Omega}(J_n) = \lambda_A \mathrm{Sign}(\widetilde{A})$. This entails that $\mathcal{P}_{\Omega}(\Delta) = 0$ and hence **(C1)** is satisfied.

Similarly, by our construction, we have $P_{\mathcal{T}}(\widetilde{Z} - \Delta) = \mathcal{P}_{\mathcal{T}}(\lambda_L \widetilde{U}\widetilde{V}^T)$ and by optimality of $(\widetilde{A}, \widetilde{L})$, we have $\mathcal{P}_{\widetilde{\mathcal{T}}}(J_n) = \mathcal{P}_{\widetilde{\mathcal{T}}}(\widetilde{Z} - \Delta)$ which by Lemma 35 entails that $\mathcal{P}_{\mathcal{T}}(J_n) = \mathcal{P}_{\mathcal{T}}(\lambda_L \widetilde{U}\widetilde{V}^T)$. Consequently, $\mathcal{P}_{\mathcal{T}}(\Delta) = 0$ and hence **(C3)** is also satisfied, considering the restriction on the oracle problem.

It suffices to show that **(C2)** and **(C4)** are satisfied with high probability. This has been shown in the next Lemma.

Lemma 34. *Under assumptions* **(A1)-(A5)**, $\widetilde{Z}$ *satisfies conditions* **(C2)** *and* **(C4)** *with probability* $1 - c_1 \exp(-c_2 n)$ *for some positive constants* c_1 *and* c_2.

This concludes the proof of the theorem.



(a) Effect of η

(b) Effect of r

(c) Effect of s

Figure 5.1: Probability of success in recovering the true signed support of A^* versus the control parameter Θ (rescaled ηn) with $p = 200$, $r = 10$ and $s = 20$ for different values of η (left), and, with $p = 200$, $s = 20$ and $\eta = 0.01$ for different number of latent time series r (middle), and, with $p = 200$, $r = 10$ and fixed $\eta = 0.01$ for different sparsity sizes s (right).

5.5 Experimental Results

5.5.1 Synthetic Data

Motivated by the example discussed in the paper, we simulate a similar (but different) dynamic system for the purpose of our experiments. Consider the system where each latent variable is only evolves by itself, i.e., $C^* = 0$ and D^* is a diagonal matrix. Moreover, assume that each latent variable affects exactly two observed variable and each observed variable is affected by exactly two latent variable, i.e., each column of B^* has $2p/r$ non-zeros and each row of B^* has two non-zeros. We randomly select a support of size s per row for A^* and draw all the values of A^* and B^* i.i.d. standard Gaussian. To make the matrix $\mathcal{A}^*$ negative definite (hence, stable), using Geršgorin disk theorem [54], we put a large-enough negative value on the diagonals of A^* and D^*.

We generate the data according to the continuous time model. The

solution to the first order system can be written as

$$\begin{bmatrix} x(t) \\ u(t) \end{bmatrix} = e^{\mathcal{A}^*(t-t_0)} \begin{bmatrix} x(t_0) \\ u(t_0) \end{bmatrix} + \int_{t_0}^{t} e^{\mathcal{A}^*(t-\tau)} dw(\tau),$$

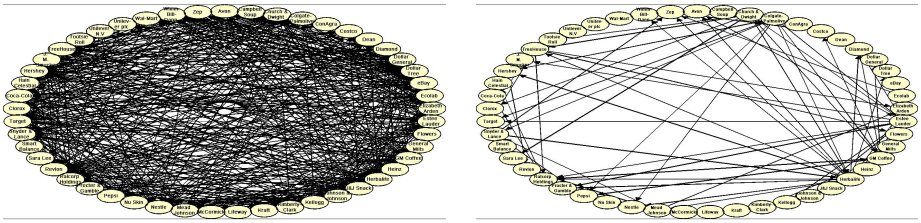
where, $e^{\mathcal{A}^*} = I + \mathcal{A}^* + \frac{1}{2}\mathcal{A}^{*2} + \dots$ is a generalization of the exponential function to matrices. In particular, if $\mathcal{A}^*$ is stable and has a singular value decomposition $\mathcal{U}^*\Sigma^*\mathcal{V}^*$, then, $e^{\mathcal{A}^*} = \mathcal{U}^* e^{\Sigma^*} \mathcal{V}^*$, where e^{Σ^*} is element-wise exponential function. We sub-sample this system at points $t_i = \eta i$ for $i = 1, 2, \dots, n$, that is

$$\begin{bmatrix} x(i) \\ u(i) \end{bmatrix} = e^{\eta\mathcal{A}} \begin{bmatrix} x(i-1) \\ u(i-1) \end{bmatrix} + \int_{\eta(i-1)}^{\eta i} e^{\mathcal{A}(\eta i - \tau)} dw(\tau)$$

The stochastic integral can be estimated by binning the interval and assuming the Brownian motion is constant over the bin and hence, can be estimated by a standard Gaussian. For more information on this integration method, we refer to Chapter 4 of [122].

Using this data, we solve (5.4) using accelerated proximal gradient method [86]. Motivated by our Theorem, we plot our result with respect to the control parameter $\Theta = \frac{\eta n}{s^3 \log((s+2r)p+r^2)}$. We pick the values of λ_A and λ_L by dividing the training data into chunks each having consecutive samples and do the cross validation over those chunks. Note that this is different from the standard cross validation technique due to the dependency of samples.

Figure 5.4.1 shows the phase transition of the probability of success in recovering the exact sign support of the matrix A^*. We ran three different experiments, each investigating the effect of one of the three key parameters of the system η (sampling frequency), r (number of latent variables) and s (sparsity of the model). These three figures show that the probability of success curves line up if they are plotted versus the correct control parameter. The first two curves for η and r line up versus Θ, indicating that our theorem suggests the correct scaling law for the sample complexity. However, from this experiment, it seems that the phase transition probability scales with s^2 not s^3. Perhaps the result of our theorem and also [15] (for $r = 0$) can be tightened.



(a) Pure LASSO (b) Our Algorithm

Figure 5.2: Comparison of the stock dependencies recovered by Pure LASSO [15] and our algorithm.

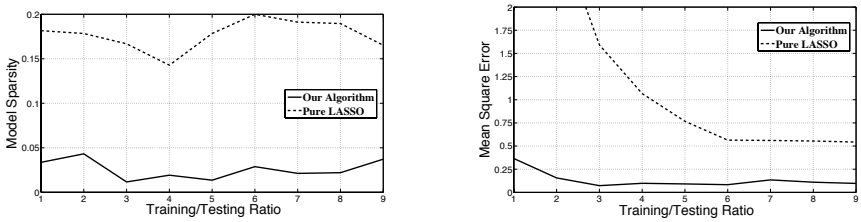
5.5.2 Stock Market Data

We take the end-of-the-day closing stock prices for 50 different companies in the period of May 17, 2010 - May 13, 2011 (255 business days). These companies (among them, Amazon, eBay, Pepsi, etc) are consumer goods companies traded either at NASDAQ or NYSE in USD. The data is collected from Google Finance website. Our goal is to observe the stock prices for a period of time and predict it for the entire days of the next month with small error.

Applying our method and pure LASSO [15] to the data, we recover the structure of the dependencies among stocks. We represent the result as a graph in Fig 5.5.2; where each company is a node in this graph and there is an edge between company i and j if $\hat{A}_{ij} \neq 0$. This result shows that the recovered dependency structure by our algorithm is order of magnitude sparser than the one recovered by pure LASSO.

To show the usefulness of our algorithm for prediction purposes, we apply our algorithm to this data and try to learn the model using the data for n (consecutive) days and then compute the mean squared error in the prediction of the following month (25 business days). We randomly pick an starting day n_0 between day 1 and day $255 - 25 - n$. Then we learn the model using the data from the day n_0 to the day $n_0 + n$ (total of n days). Then, we test our data on the consecutive 25 days. Finally, we average the error over 10 different starting points n_0 for each value of n. We pick the regularizers by the semi-cross validation process explained in the previous section. The ratio $\frac{n}{25}$ shows the ratio of training sample size to the testing sample size.

Figure 5.3(b) shows the prediction error for both our method and pure



(a) Model Sparsity (b) Prediction Error

Figure 5.3: Prediction error and model sparsity versus the ratio of the training/testing sample sizes for prediction of the stock price. Prediction error is measured using mean squared error and the model sparsity is the number of non-zero entries divided by the size of $\widehat{A}$.

LASSO [15] method as the train/test ratio increases. It can be seen that our method not only have better prediction, but also is more robust. Our algorithm requires only 3 months of the past data to give a robust estimation of the next month; in contrast with almost 6 months requirement of LASSO. However, the error of our algorithm is much smaller (by a factor of 6) than LASSO even in the steady state. Figure 5.3(a) shows the sparsity level for our model and the LASSO model. The number of latent variables our model finds varies from $8-12$ for different train/test ratios. As Figure 5.3(a) illustrates, our estimated $\widehat{A}$ is order of magnitude sparser than the one estimated by LASSO.

5.6 Proof of Lemma 32

Proof. Ignoring the term $\|x(n+1)-x(n)\|_2^2$ which is independent of A, minimization of $\mathcal{L}(A)$ with this infinite sample size is equivalent to

$$\begin{aligned}
&\min_A \mathbb{E}\left[x(n)^T A^T A x(n) - \frac{2}{\eta}(x(n+1)-x(n))^T A x(n)\right] \\
&= \min_A \mathbb{E}\left[\mathbf{Trace}\left(Ax(n)x(n)^T A^T\right) - 2\left(A^* x(n) + B^* u(t)\right)^T A x(n)\right] \\
&= \min_A \mathbf{Trace}\left(AQ^*A^T\right) - 2\mathbf{Trace}\left(A^*Q^*A^T\right) - 2\mathbf{Trace}\left(B^*R^*A^T\right) \\
&= \min_A \mathbf{Trace}\left(\left(A - 2\left(A^* + B^*R^*(Q^*)^{-1}\right)\right)Q^*A^T\right).
\end{aligned}$$

Here we ignored the term $w(n)$ due to the fact that it is zero mean and independent of $x(n)$ and $u(n)$. This implies that the asympotatic optimizer of $\mathcal{L}(\cdot)$ satisfies $\hat{A} = A^* + B^* R^* (Q^*)^{-1}$. □

5.7 Illustrative Example

In this section, we analyze the illustrative example discussed in Sec 5.3. For that example, Lyapunov stability equation requires

$$\begin{bmatrix} -2Q^* + B^*R^* + R^{*T}B^{*T} & -2R^{*T} + B^* \\ -2R^* + B^{*T} & -2P^* \end{bmatrix} = \begin{bmatrix} -I & 0 \\ 0 & -I \end{bmatrix}.$$

This entails that $R^* = \frac{1}{2}B^{*T}$ and $Q^* = \frac{1}{2}(I + B^*B^{*T})$ with $C_{\min} = \frac{1}{2}$. It can be easily checked that $Q^{*-1} = 2(I - \frac{r}{p+r}B^*B^{*T})$. Thus, the low-rank matrix of interest is

$$\begin{aligned} L^* &= B^*R^*Q^{*-1} \\ &= B^*B^{*T}(I - \frac{r}{p+r}B^*B^{*T}) \\ &= (1 - \frac{p}{p+r})B^*B^{*T}. \end{aligned}$$

Taking singular value decomposition $U^*\Sigma^*V^*$ of this matrix, we get $U^* = V^* = \sqrt{\frac{r}{p}}B^*$ and hence $\mu = r$. Considering $s = 1$, the identifiability assumption **(A2)** becomes $\alpha = \frac{3r}{\sqrt{p}} \leq 1$ or equivalently, $r \leq \frac{\sqrt{p}}{3}$.

Considering assumption **(A3)**, note that $Q^*_{\mathcal{S}_k\mathcal{S}_k} = 1$ is just an scalar since $s = 1$. Moreover, $Q^*_{\mathcal{S}_{k^c}\mathcal{S}_k}$ is a vector with all entries equal to $\frac{1}{2}$ and hence $\theta = \frac{1}{2}$.

5.8 Proof of Lemma 33

Proof. Suppose $\widetilde{A} = \widehat{A} + D_A$ and $\widetilde{L} = \widehat{L} + D_L$ for some matrices D_A and D_L. From condition (C5), we have

$$\mathcal{L}(\widehat{A} + \widehat{L}) \geq \mathcal{L}(\widetilde{A} + \widetilde{L}) - \left\langle -\widetilde{Z}, D_A + D_L \right\rangle.$$

Let $Z_A = \lambda_A \text{Sign}(\widetilde{A}) - F$ with $\mathcal{P}_\Omega(F) = 0$ and $\langle F, D_A \rangle = \lambda_A \|\mathcal{P}_{\Omega^c}(D_A)\|_1$.

By zero duality gap between dual norms, F exists and hence, Z_A is in the subgradient of $\lambda_A\|\widetilde{A}\|_1$, i.e.,

$$\lambda_A\|\widehat{A}\|_1 \geq \lambda_A\|\widetilde{A}\|_1 - \langle Z_A, D_A\rangle\,.$$

Let $\widetilde{\mathcal{T}} = \mathcal{T}(\widetilde{L})$ with SVD decomposition $\widetilde{L} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$. Now, let $Z_L = \lambda_L U^* V^{*T} + W_1 - W_0$ with $\mathcal{P}_{\mathcal{T}}(W_0) = 0$ and $\|W_0\|_2 \leq (1-\alpha)\lambda_L$ and $\langle W_0, D_L\rangle = (1-\alpha)\lambda_L\|\mathcal{P}_{\mathcal{T}^c}(D_L)\|_*$ and with $\mathcal{P}_{\mathcal{T}^c}(W_1) = 0$ and $\|W_1\|_2 \leq 4\rho\lambda_L$. Lemma 35 ensures that $Z_L = \lambda_L\widetilde{U}\widetilde{V}^T + W_2$ for some matrix W_2 with $\mathcal{P}_{\widetilde{\mathcal{T}}^c}(W_2) = 0$ and $\|W_2\|_2 < \lambda_L$ and hence, Z_L is in the subgradient of $\lambda_L\|\widetilde{L}\|_*$, i.e.,

$$\lambda_L\|\widehat{L}\|_1 \geq \lambda_L\|\widetilde{L}\|_1 - \langle Z_L, D_L\rangle\,.$$

Combining these three inequalities, we get

$$\begin{aligned}\mathcal{L}(\widehat{A}+\widehat{L}) + \lambda_A\|\widehat{A}\|_1 + \lambda_L\|\widehat{L}\|_* \geq \mathcal{L}(\widetilde{A}+\widetilde{L}) + \lambda_A\|\widetilde{A}\|_1 + \lambda_L\|\widetilde{L}\|_* \\ + \left\langle \widetilde{Z}, D_A + D_L\right\rangle - \langle Z_A, D_A\rangle - \langle Z_L, D_L\rangle\,.\end{aligned}$$

It suffices to show that $\left\langle \widetilde{Z}, D_A + D_L\right\rangle - \langle Z_A, D_A\rangle - \langle Z_L, D_L\rangle \geq 0$ to conclude by the optimality of $(\widehat{A}, \widehat{L})$ that the result holds. Notice that (C5) ensures that $\mathcal{P}_\Omega(\widetilde{Z}) = \lambda_A \mathrm{Sign}(\widetilde{A})$ and $\mathcal{P}_{\widetilde{\mathcal{T}}}(\widetilde{Z}) = \lambda_L\widetilde{U}\widetilde{V}^T$. By Lemma 35, we conclude that $\mathcal{P}_{\mathcal{T}}(\widetilde{Z}) = \lambda_L U^* V^{*T} + W_1$. Thus, for some $\gamma < 1$, we have

$$
\begin{aligned}
&\left\langle \widetilde{Z}, D_A + D_L \right\rangle - \langle Z_A, D_A \rangle - \langle Z_L, D_L \rangle \\
&\quad = \lambda_A \|\mathcal{P}_{\Omega^c}(D_A)\|_1 + (1-\alpha)\lambda_L \|\mathcal{P}_{\mathcal{T}^c}(D_L)\|_* \\
&\qquad + \left\langle \widetilde{Z}, D_A + D_L \right\rangle - \left\langle \lambda_A \mathrm{Sign}(\widetilde{A}), \mathcal{P}_\Omega(D_A) \right\rangle \\
&\qquad - \left\langle \lambda_L U^* V^{*T} + W_1, \mathcal{P}_{\mathcal{T}}(D_L) \right\rangle \\
&\quad = \lambda_A \|\mathcal{P}_{\Omega^c}(D_A)\|_1 + (1-\alpha)\lambda_L \|\mathcal{P}_{\mathcal{T}^c}(D_L)\|_* \\
&\qquad + \underbrace{\left\langle \mathcal{P}_{\Omega^c}(\widetilde{Z}), \mathcal{P}_{\Omega^c}(D_A) \right\rangle}_{\geq -\gamma\lambda_A \|\mathcal{P}_{\Omega^c}(D_A)\|_1 \quad by\ (C2)} + \underbrace{\left\langle \mathcal{P}_{\mathcal{T}^c}(\widetilde{Z}), \mathcal{P}_{\mathcal{T}^c}(D_L) \right\rangle}_{\geq -\gamma(1-\alpha)\lambda_L \|\mathcal{P}_{\mathcal{T}^c}(D_L)\|_* \quad by\ (C4)} \\
&\qquad + \underbrace{\left\langle \mathcal{P}_{\mathcal{T}}(\widetilde{Z}) - \lambda_L U^* V^{*T} - W_1, \mathcal{P}_{\mathcal{T}}(D_L) \right\rangle}_{=0 \quad by\ (C3)} \\
&\qquad + \underbrace{\left\langle \mathcal{P}_{\Omega}(\widetilde{Z}) - \lambda_A \mathrm{Sign}(\widetilde{A}), \mathcal{P}_{\Omega}(D_A) \right\rangle}_{=0 \quad by\ (C1)} \\
&\quad \geq (1-\gamma)\lambda_A \|\mathcal{P}_{\Omega^c}(D_A)\|_1 + (1-\gamma)(1-\alpha)\lambda_L \|\mathcal{P}_{\mathcal{T}^c}(D_L)\|_* \geq 0.
\end{aligned}
$$

This concludes the proof of the lemma.

□

Lemma 35. *For any two matrices with SVD $L^* = U^*\Sigma^*V^{*T}$ and $\widetilde{L} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$ and corresponding tanget spaces $\mathcal{T}$ and $\widetilde{\mathcal{T}}$, if $\rho(\mathcal{T}, \widetilde{\mathcal{T}}) = \rho \leq \frac{\alpha}{4}$, then, for any matrix W_0 with $\mathcal{P}_{\mathcal{T}}(W_0) = 0$ and $\|W_0\|_2 \leq (1-\alpha)\lambda_L$, there exist matrices W_1 and W_2 such that*

$$\widetilde{Z}_L := \lambda_L U^* V^{*T} + W_1 - W_0 = \lambda_L \widetilde{U}\widetilde{V}^T + W_2,$$

where, $\mathcal{P}_{\widetilde{\mathcal{T}}}(W_2) = 0$ with $\|W_2\|_2 < \lambda_L$ and $\mathcal{P}_{\mathcal{T}^c}(W_1) = 0$ with $\|W_1\|_2 \leq 4\rho\lambda_L$.

Proof. Let $W_2 = -\mathcal{P}_{\mathcal{T}^c}(\lambda_L \widetilde{U}\widetilde{V}^T) - W_0$ and $W_1 = \mathcal{P}_{\mathcal{T}}(\lambda_L \widetilde{U}\widetilde{V}^T) - \lambda_L U^* V^{*T}$. For this choice, the equality constraints hold. First, notice that $\|\mathcal{P}_{\mathcal{T}}(M)\|_2 \leq 2\|M\|_2$ and hence,

$$
\begin{aligned}
\left\|\widetilde{Z}_L\right\|_2 &= \left\|\lambda_L \widetilde{U}\widetilde{V}^T - \mathcal{P}_{\mathcal{T}^c}(\lambda_L \widetilde{U}\widetilde{V}^T) - W_0\right\|_2 \\
&\leq \left\|\mathcal{P}_{\mathcal{T}}(\lambda_L \widetilde{U}\widetilde{V}^T)\right\|_2 + \|W_0\|_2 \\
&\leq 2\left\|\lambda_L \widetilde{U}\widetilde{V}^T\right\|_2 + \|W_0\|_2 \leq (3-\alpha)\lambda_L.
\end{aligned}
$$

Using this, we can bound both W_1 and W_2. For W_2, we have

$$\begin{aligned}
\|W_2\|_2 &\leq \left\|\mathcal{P}_{\mathcal{T}^c}(\lambda_L \widetilde{U}\widetilde{V}^T)\right\|_2 + \|W_0\|_2 \\
&= \left\|\mathcal{P}_{\mathcal{T}^c}(\lambda_L \widetilde{U}\widetilde{V}^T - \lambda_L U^* V^{*T} - W_1)\right\|_2 + \|W_0\|_2 \\
&\leq \left\|\lambda_L \widetilde{U}\widetilde{V}^T - \lambda_L U^* V^{*T} - W_1\right\|_2 + \|W_0\|_2 \\
&= \left\|\mathcal{P}_{\widetilde{\mathcal{T}}}(\widetilde{Z}_L) - \mathcal{P}_{\mathcal{T}}(\widetilde{Z}_L)\right\|_2 + \|W_0\|_2 \\
&\leq \rho \left\|\widetilde{Z}_L\right\|_2 + \|W_0\|_2 \\
&\leq ((3-\alpha)\rho + (1-\alpha))\,\lambda_L < \lambda_L.
\end{aligned}$$

Note that $\mathcal{P}_{\widetilde{\mathcal{T}}}(\widetilde{Z}_L) = \lambda_L \widetilde{U}\widetilde{V}^T$ and hence, we can establish

$$\begin{aligned}
\|W_1\|_2 &= \left\|\mathcal{P}_{\mathcal{T}}(\lambda_L \widetilde{U}\widetilde{V}^T) - \lambda_L U^* V^{*T}\right\|_2 \\
&= \left\|\mathcal{P}_{\mathcal{T}}(\widetilde{Z}_L) - \mathcal{P}_{\widetilde{\mathcal{T}}}(\widetilde{Z}_L)\right\|_2 + \left\|\lambda_L \widetilde{U}\widetilde{V}^T - \lambda_L U^* V^{*T}\right\|_2 \\
&\leq \rho \left\|\widetilde{Z}_L\right\|_2 + \rho\lambda_L \leq ((3-\alpha)\rho + \rho)\,\lambda_L.
\end{aligned}$$

This concludes the proof of the lemma.

□

5.9 Auxiliary Optimality Lemmas

General Notation: For a matrix $X \in \mathbb{R}^{a\times b}$, we use $X^{(1)}, \ldots, X^{(a)}$ to denote rows, $X_1, \ldots, X_b$ to denote columns and $X_1^{(1)}, \ldots, X_b^{(a)}$ to denote entries. Also, for the sets of indecies $\mathcal{S}_1 \subseteq \{1, \cdots, a\}$ and $\mathcal{S}_2 \subseteq \{1, \cdots, b\}$, the matrix $X_{\mathcal{S}_1\mathcal{S}_2} \in \mathbb{R}^{|\mathcal{S}_1|\times|\mathcal{S}_2|}$ represents the sub-matrix of X consisting of the rows and columns corresponding to index sets $\mathcal{S}_1$ and $\mathcal{S}_2$.

Lemma 36 (Convex Optimality)**.** *If $\widehat{A}$ is a solution of* (5.4) *then there exists a matrix $\widehat{Z} \in \mathbb{R}^{p\times p}$, called* dual variable*, such that $\widehat{Z} \in \lambda_A \partial\|\widehat{A}\|_1$ and $\widehat{Z} \in \lambda_L \partial\|\widehat{L}\|_*$ and*

$$-\frac{1}{\eta n}\sum_{i=1}^{n}\Big(x(i+1) - x(i) - \eta(\widehat{A}+\widehat{L})x(i)\Big)\,x(i)^T + \widehat{Z} = 0. \tag{5.6}$$

Proof. The proof follows from the standard first order optimality argument. □

Lemma 37. *For constructed dual variable, conditions (C2) and (C4) are satisfied with high probability.*

Proof. Let $Q^{(n)} = \frac{1}{n}\sum_{i=1}^{n} x(i)x(i)^T$ and $R^{(n)} = \frac{1}{n}\sum_{i=1}^{n} u(i)x(i)^T$. Substituting $x(i+1) - x(i) = \eta A^* x(i) + \eta B^* u(i) + w(i)$ and $L^* = B^* R^* (Q^*)^{-1}$ in (C5), we equivalently get

$$\begin{aligned}(\widetilde{A} - A^*)Q^{(n)} + (\widetilde{L} - L^*)Q^{(n)} - \underbrace{B^*\big(R^{(n)} - R^*(Q^*)^{-1}Q^{(n)}\big)}_{Y^{(n)}} \\ - \underbrace{\frac{1}{n\eta}\sum_{i=1}^{n} w(i)x(i)^T}_{W^{(n)}} + \widetilde{Z} = 0.\end{aligned} \tag{5.7}$$

We can rewrite this equation as

$$\begin{aligned}\mathcal{P}_{\Omega^c}(\widetilde{L} - L^*)Q^{(n)} + (\widetilde{A} - A^* + \mathcal{P}_{\Omega}(\widetilde{L} - L^*))Q^{(n)} \\ - Y^{(n)} - W^{(n)} + \widetilde{Z} = 0.\end{aligned} \tag{5.8}$$

Let us only focus on the k^{th} row of the system of equation (5.7). We can break down (5.7) on the k^{th} row into two sets of linear equations as follows:

$$\begin{aligned}&(\widetilde{A} - A^* + \widetilde{L} - L^*)^{(k)}_{\mathcal{S}_k} Q^{(n)}_{\mathcal{S}_k \mathcal{S}_k} \\ &\qquad = -(\widetilde{L} - L^*)^{(k)}_{\mathcal{S}_k^c} Q^{(n)}_{\mathcal{S}_k^c \mathcal{S}_k} + Y^{(n)}_{\mathcal{S}_k} + W^{(n)}_{\mathcal{S}_k} - \widetilde{Z}_{\mathcal{S}_k} \\ &(\widetilde{A} - A^* + \widetilde{L} - L^*)^{(k)}_{\mathcal{S}_k} Q^{(n)}_{\mathcal{S}_k \mathcal{S}_k^c} \\ &\qquad = -(\widetilde{L} - L^*)^{(k)}_{\mathcal{S}_k^c} Q^{(n)}_{\mathcal{S}_k^c \mathcal{S}_k^c} + Y^{(n)}_{\mathcal{S}_k^c} + W^{(n)}_{\mathcal{S}_k^c} - \widetilde{Z}_{\mathcal{S}_k^c}.\end{aligned} \tag{5.9}$$

By Lemma 38, we have

$$\begin{aligned}&\left\| (\widetilde{L} - L^*)^{(k)}_{\mathcal{S}_k^c} Q^{(n)}_{\mathcal{S}_k^c \mathcal{S}_k} \left(Q^{(n)}_{\mathcal{S}_k \mathcal{S}_k}\right)^{-1} \right\|_\infty \\ &\qquad \leq \left(1 - \frac{\theta}{2}\right) \left\| \mathcal{P}_{\Omega^c}(\widetilde{L} - L^*) \right\|_\infty \leq \left\| \widetilde{L} - L^* \right\|_\infty .\end{aligned}$$

Since $\widetilde{Z}$ satisfies (C3), we have $\left\|\mathcal{P}_{\mathcal{T}}(\widetilde{U}\widetilde{V}^T) - U^*V^{*T}\right\|_2 \leq 4\rho$. By the properties of the oracle problem (closeness of spaces $\mathcal{T}$ and $\widetilde{\mathcal{T}}$), we have

$$\begin{aligned}
&\left\|\widetilde{L} - L^*\right\|_2 \\
&\quad\leq \left\|\mathcal{P}_{\widetilde{\mathcal{T}}}(\widetilde{L} - L^*) - \mathcal{P}_{\mathcal{T}}(\widetilde{L} - L^*)\right\|_2 \\
&\qquad + \left\|\mathcal{P}_{\widetilde{\mathcal{T}}}(L^*) - \mathcal{P}_{\mathcal{T}}(L^*)\right\|_2 + \left\|\mathcal{P}_{\mathcal{T}}(\widetilde{L}) - L^*\right\|_2 \\
&\quad\leq \rho\left\|\widetilde{L} - L^*\right\|_2 + \rho\left\|L^*\right\|_2 \\
&\qquad + \left\|\mathcal{P}_{\mathcal{T}}(\widetilde{U}\widetilde{V}^T) - U^*V^{*T}\right\|_2 \left\|\widetilde{L} - L^*\right\|_2 \\
&\quad\leq \rho\left\|\widetilde{L} - L^*\right\|_2 + \rho\left\|L^*\right\|_2 + 4\rho\left\|\widetilde{L} - L^*\right\|_2 .
\end{aligned}$$

Hence,

$$\left\|\widetilde{L} - L^*\right\|_\infty \leq \left\|\widetilde{L} - L^*\right\|_2 \leq \frac{\rho}{1-5\rho}\left\|L^*\right\|_2 . \tag{5.10}$$

Thus, from the first equation in (5.9) and Lemma 39, we get

$$\begin{aligned}
&\left\|\widetilde{A} - A^*\right\|_\infty \\
&\quad\leq \frac{2\rho}{1-5\rho}\left\|L^*\right\|_2 + \frac{\sqrt{s}}{\mathcal{C}_{\min}}\left(\left\|Y^{(n)}\right\|_\infty + \left\|W^{(n)}\right\|_\infty + \lambda_A\right) \\
&\quad\leq \frac{2\rho}{1-5\rho}\left\|L^*\right\|_2 + \frac{(8-\theta)\lambda_A\sqrt{s}}{\mathcal{C}_{\min}(4-\theta)} \\
&\quad\leq \left(\frac{\alpha\theta}{\mathcal{D}_{\max}\left(1 + \frac{\mathcal{D}_{\max}}{\mathcal{C}_{\min}}\right)} + \frac{(8-\theta)\sqrt{s}}{\mathcal{C}_{\min}(4-\theta)}\right)\lambda_A .
\end{aligned} \tag{5.11}$$

The last inequality follows from Lemmas 40 and 41. Substituting $(\widetilde{A} - A^* + \widetilde{L} - L^*)^{(k)}_{\mathbb{S}_k}$ from the first equation in the second in (5.9), we get

$$\begin{aligned}
\widetilde{Z}_{\mathbb{S}_k^c} &= -(\widetilde{L} - L^*)^{(k)}_{\mathbb{S}_k^c} Q^{(n)}_{\mathbb{S}_k^c\mathbb{S}_k^c} + Y^{(n)}_{\mathbb{S}_k^c} + W^{(n)}_{\mathbb{S}_k^c} \\
&\quad - \left(-(\widetilde{L} - L^*)^{(k)}_{\mathbb{S}_k^c} Q^{(n)}_{\mathbb{S}_k^c\mathbb{S}_k} + Y^{(n)}_{\mathbb{S}_k} + W^{(n)}_{\mathbb{S}_k} - \widetilde{Z}_{\mathbb{S}_k}\right)\left(Q^{(n)}_{\mathbb{S}_k\mathbb{S}_k}\right)^{-1} Q^{(n)}_{\mathbb{S}_k\mathbb{S}_k^c} .
\end{aligned}$$

Taking maximum absolute value from both sides, using Lemmas 38,40 and 41, we get

$$
\begin{aligned}
&\left\|\mathcal{P}_{\Omega^c}(\widetilde{Z})\right\|_\infty \\
&\leq \max_k \left\|(\widetilde{L}-L^*)^{(k)}_{\mathcal{S}_k^c}\left(Q^{(n)}_{\mathcal{S}_k^c\mathcal{S}_k^c}-Q^{(n)}_{\mathcal{S}_k^c\mathcal{S}_k}\left(Q^{(n)}_{\mathcal{S}_k\mathcal{S}_k}\right)^{-1}Q^{(n)}_{\mathcal{S}_k\mathcal{S}_k^c}\right)\right\|_\infty \\
&\quad + \left\|Y^{(n)}\right\|_\infty + \left\|W^{(n)}\right\|_\infty \\
&\quad + \max_k \left\|Q^{(n)}_{\mathcal{S}_k^c\mathcal{S}_k}\left(Q^{(n)}_{\mathcal{S}_k\mathcal{S}_k}\right)^{-1}\right\|_{\infty,1}\left(\left\|Y^{(n)}\right\|_\infty + \left\|W^{(n)}\right\|_\infty + \lambda_A\right) \\
&\leq \max_k \left\|(\widetilde{L}-L^*)^{(k)}_{\mathcal{S}_k^c}\left(Q^{(n)}_{\mathcal{S}_k^c\mathcal{S}_k^c}-Q^{(n)}_{\mathcal{S}_k^c\mathcal{S}_k}\left(Q^{(n)}_{\mathcal{S}_k\mathcal{S}_k}\right)^{-1}Q^{(n)}_{\mathcal{S}_k\mathcal{S}_k^c}\right)\right\|_\infty \\
&\quad + \frac{\theta\lambda_A}{4(4-\theta)} + \frac{\theta\lambda_A}{4(4-\theta)} \\
&\quad + \left(1-\frac{\theta}{2}\right)\left(\frac{\theta\lambda_A}{4(4-\theta)} + \frac{\theta\lambda_A}{4(4-\theta)} + \lambda_A\right) \\
&\leq \frac{2\rho}{1-5\rho}\left(1+\frac{\mathcal{D}_{\max}}{\mathcal{C}_{\min}}\right)\mathcal{D}_{\max}\left\|L^*\right\|_2 + \left(1-\frac{\theta}{4}\right)\lambda_A \\
&\leq \left(1-\frac{(1-\alpha)\theta}{4}\right)\lambda_A.
\end{aligned}
$$

The one to the last inequality follows from perturbation theory for the Schur complement of semi-definite matrices (See Lemma 44). The last inequality holds for our choice of ρ. Hence, condition (C2) is satisfied.

To show (C3) also holds, notice that from (5.8), we have

$$
\begin{aligned}
&\left\|\mathcal{P}_{\mathcal{T}}(\widetilde{Z})\right\|_2 \\
&\quad \leq \left\|\mathcal{P}_{\mathcal{T}^c}\left((\widetilde{A}+\widetilde{L}-A^*-L^*)Q^{(n)}\right)\right\|_2 + \left\|Y^{(n)}\right\|_2 + \left\|W^{(n)}\right\|_2 \\
&\quad \leq \left\|\mathcal{P}_{\mathcal{T}^c}\left((\widetilde{A}+\widetilde{L}-A^*-L^*)Q^{(n)}\right)\right\|_2 + \frac{\theta\lambda_A\sqrt{p}}{2(4-\theta)}.
\end{aligned}
$$

The last inequality follows from Lemmas 40 and 41 and the fact that $Q^{(n)}$ on the support is invertible for the given sample complexity due to Lemma 39.

Next, notice that $L^* = B^*R^*(Q^*)^{-1}$ and hence the row-space of L^* is the column/row space of Q^* and consequently, for any matrix $F \in \mathcal{T}$, we have

$\mathcal{P}_{\mathcal{T}^c}(FQ^*) = 0$. Thus, we have

$$\begin{aligned}
&\left\|\mathcal{P}_{\mathcal{T}^c}\left((\widetilde{A}+\widetilde{L}-A^*-L^*)Q^{(n)}\right)\right\|_2 \\
&\qquad = \left\|\mathcal{P}_{\mathcal{T}^c}\left((\widetilde{A}+\widetilde{L}-A^*-L^*)\left(Q^{(n)}-Q^*\right)\right)\right\|_2 \\
&\qquad\qquad + \left\|\mathcal{P}_{\mathcal{T}^c}\left(\widetilde{A}+\widetilde{L}-A^*-L^*\right)Q^*\right\|_2 \\
&\qquad \leq \left\|(\widetilde{A}+\widetilde{L}-A^*-L^*)\left(Q^{(n)}-Q^*\right)\right\|_2 \\
&\qquad\qquad + \left\|\widetilde{A}+\widetilde{L}-A^*-L^*\right\|_2 \|Q^*\|_2 \sqrt{p} \\
&\qquad \leq \left(\sqrt{s}\left\|\widetilde{A}-A^*\right\|_\infty + \left\|\widetilde{L}-L^*\right\|_2\right) \\
&\qquad\qquad \left(\sqrt{p}\left\|Q^{(n)}-Q^*\right\|_\infty + \mathcal{D}_{\max}\right)\sqrt{p}.
\end{aligned}$$

Finally, from (5.11), (5.10) and Lemma 43, we get

$$\begin{aligned}
&\left\|\mathcal{P}_{\mathcal{T}^c}(\widetilde{Z})\right\|_2 \\
&\leq \left(\frac{\theta \mathcal{C}_{\min}\sqrt{p}}{9s\sqrt{s}} + \mathcal{D}_{\max}\right) \\
&\qquad \left(\frac{3\alpha\theta\sqrt{s}}{2\mathcal{D}_{\max}\left(1+\frac{\mathcal{D}_{\max}}{\mathcal{C}_{\min}}\right)} + \frac{(8-\theta)s}{\mathcal{C}_{\min}(4-\theta)}\right)\lambda_A\sqrt{p} + \frac{\theta\lambda_A\sqrt{p}}{2(4-\theta)} \\
&\leq \theta(1-\alpha)\lambda_L.
\end{aligned}$$

Hence, condition (C4) is also satisfied. This concludes the proof of the lemma. □

5.10 Concentration Results

In this section we prove the concentration results used throughout the paper. Before, we state the results, we want to introduce some useful notations and inequalities used to get the results. By the dynamics of the system, we have

$$\begin{bmatrix} x(i) \\ u(i) \end{bmatrix} = (I+\eta\mathcal{A}^*)^i \begin{bmatrix} x(0) \\ u(0) \end{bmatrix} + \sum_{l=0}^{i-1} (I+\eta\mathcal{A}^*)^{i-l-1} w(l).$$

Lemma 38. *For any $\mathcal{S} \subseteq \{1, 2, \ldots, p\}$ with $|\mathcal{S}| \leq s$ and sample complexity $n\eta \geq \frac{3\times 10^6\, s^3}{D^2\,\theta^2\,\mathcal{C}_{\min}^2} \log\left(\frac{4((s+2r)p+r^2)}{\delta}\right)$ with high probability, we have*

$$\left\| \mathcal{Q}^{(n)}_{\mathcal{S}^c\mathcal{S}} \left(\mathcal{Q}^{(n)}_{\mathcal{S}\mathcal{S}}\right)^{-1} \right\|_{\infty,1} \leq 1 - \frac{\theta}{2},$$

provided that $\left\|\mathcal{Q}^*_{\mathcal{S}^c\mathcal{S}} \left(\mathcal{Q}^*_{\mathcal{S}\mathcal{S}}\right)^{-1}\right\|_{\infty,1} \leq 1 - \theta$.

Proof. Using Lemma 39, it can be shown (see Lemma A.1 in [15] for example) that

$$\begin{aligned}\left\| \mathcal{Q}^{(n)}_{\mathcal{S}^c\mathcal{S}} \left(\mathcal{Q}^{(n)}_{\mathcal{S}\mathcal{S}}\right)^{-1} \right\|_{\infty,1} \leq{}& \left\|\mathcal{Q}^*_{\mathcal{S}^c\mathcal{S}} \left(\mathcal{Q}^*_{\mathcal{S}\mathcal{S}}\right)^{-1}\right\|_{\infty,1} \\ &+ \frac{3|\mathcal{S}|\sqrt{|\mathcal{S}|}}{\mathcal{C}_{\min}} \left\|\mathcal{Q}^{(n)} - \mathcal{Q}^*\right\|_{\infty} \\ &+ \frac{2|\mathcal{S}|^2\sqrt{|\mathcal{S}|}}{\mathcal{C}^2_{\min}} \left\|\mathcal{Q}^{(n)} - \mathcal{Q}^*\right\|^2_{\infty}.\end{aligned}$$

The result follows from Lemma 42. This concludes the proof of the lemma.

□

Lemma 39. *For any $\mathcal{S} \subseteq \{1, 2, \ldots, p\}$ with $|\mathcal{S}| \leq s$ and sample complexity $n\eta \geq \frac{3\times 10^6\, s^3}{D^2\,\theta^2\,\mathcal{C}_{\min}^2} \log\left(\frac{4((s+2r)p+r^2)}{\delta}\right)$ with high probability, we have*

$$\Lambda_{\min}\left(\mathcal{Q}^{(n)}_{\mathcal{S}\mathcal{S}}\right) \geq \frac{\mathcal{C}_{\min}}{2}.$$

Proof. By the Courant-Fischer variational representation [62], we have

$$\begin{aligned}\Lambda_{\min}\left(\mathcal{Q}^{(n)}_{\mathcal{S}\mathcal{S}}\right) &\geq \Lambda_{\min}\left(\mathcal{Q}^*_{\mathcal{S}\mathcal{S}}\right) - \Lambda_{\max}\left(\mathcal{Q}^*_{\mathcal{S}\mathcal{S}} - \mathcal{Q}^{(n)}_{\mathcal{S}\mathcal{S}}\right) \\ &\geq \mathcal{C}_{\min} - \sqrt{s}\left\|\mathcal{Q}^* - \mathcal{Q}^{(n)}\right\|_{\infty}.\end{aligned}$$

The last inequality follows from Lemma 42. This concludes the proof of the lemma.

□

Lemma 40. *For*

$$\lambda_A \geq \frac{16(4-\theta)\sqrt{\|x(0)\|_2^2 + \|u(0)\|_2^2 + (\sqrt{\eta}+1)^2}}{\theta\sqrt{D}} \sqrt{\frac{\log\left(\frac{4((s+2r)p+r^2)}{\delta}\right)}{n\eta}}$$

with high probability, we have

$$\left\|\mathcal{W}^{(n)}\right\|_\infty \leq \frac{\theta\lambda_A}{4(4-\theta)}.$$

Proof. Let $X(i) = [x(i)\ u(i)]^T$. According to the dynamics of the system, we have

$$\begin{aligned}\mathcal{W}^{(n)} = &\frac{1}{\eta n}\sum_{i=0}^{n-1} w(i) \underbrace{X(0)^T\left((I+\eta\mathcal{A}^*)^i\right)^T}_{E_1(i)} \\ &+ \frac{1}{\eta n}\sum_{i=1}^{n-1} w(i) \underbrace{\sum_{l=0}^{i-1} w(l)^T\left((I+\eta\mathcal{A}^*)^{i-l-1}\right)^T}_{E_2(i)}.\end{aligned}$$

We bound these two terms separately. Notice that $w(i)$ is distributed $\mathcal{N}(0,\eta I)$ independent of $x(0)$ and $w(j)$'s. Given $x(0)$, we have

$$w(i)_j E_1(i)^{(k)} \sim \mathcal{N}\left(0, \eta\left(E_1(i)^{(k)}\right)^2\right).$$

By stability assumption, we have $\left(E_1(i)^{(k)}\right)^2 \leq (\|x(0)\|_2^2 + \|u(0)\|_2^2)\Sigma_{\max}^{2i}$ and hence,

$$\begin{aligned}&\mathrm{VAR}\left(\frac{1}{\eta n}\sum_{i=0}^{n-1} w(i)_j E_1(i)^{(k)}\right) \\ &\qquad\leq \frac{1}{\eta^2 n^2}\sum_{i=0}^{n-1}\mathrm{VAR}\left(w(i)_j E_1(i)^{(k)}\right) \\ &\qquad\leq \frac{\|x(0)\|_2^2 + \|u(0)\|_2^2}{\eta n(1-\Sigma_{\max}^2)}.\end{aligned}$$

Consequently, by standard concentration of Gaussian random variables and union bound, we get

$$\begin{aligned}
&\mathbb{P}\left[\left\|\frac{1}{\eta n}\sum_{i=0}^{n-1} w(i)E_1(i)\right\|_\infty \geq \epsilon\right] \\
&\qquad \leq \sum_{j=1}^{p}\sum_{k=1}^{p}\mathbb{P}\left[\left|\frac{1}{\eta n}\sum_{i=0}^{n-1} w(i)_j E_1(i)^{(k)}\right| \geq \epsilon\right] \\
&\qquad \leq 2\exp\Big(-\frac{\epsilon^2(1-\Sigma_{\max}^2)}{2\left(\|x(0)\|_2^2+\|u(0)\|_2^2\right)}\eta n \\
&\qquad\qquad\qquad + \log((s+2r)p+r^2)\Big).
\end{aligned}$$

With similar analysis, we get

$$\begin{aligned}
&\mathrm{VAR}\left(\frac{1}{\eta n}\sum_{i=0}^{n-1} w(i)_j E_2(i)^{(k)}\right) \\
&\qquad \leq \frac{1}{\eta^2 n^2}\sum_{i=0}^{n-1}\mathrm{VAR}\left(w(i)_j E_2(i)^{(k)}\right) \\
&\qquad \leq \frac{\left(\sqrt{\eta}+1\right)^2}{\eta n(1-\Sigma_{\max}^2)}.
\end{aligned}$$

The last inequality follows from the concentration of χ^2 random variables [80], in particular,

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{\eta n}\sum_{l=0}^{n-2}\|w(l)_j E_2(l)^{(k)}\|_2^2 \geq \frac{\left(1+\sqrt{\eta}\right)^2}{1-\Sigma_{\max}^2}\right] \\
&\qquad\qquad \leq \exp\left(-\frac{1}{2}\eta n + \log((s+2r)p+r^2)\right).
\end{aligned}$$

Finally, we get

$$\begin{aligned}
&\mathbb{P}\left[\left\|\frac{1}{\eta n}\sum_{i=0}^{n-1} w(i)E_2(i)\right\|_\infty \geq \epsilon\right] \\
&\quad\leq \sum_{j=1}^{p}\sum_{k=1}^{p}\mathbb{P}\left[\left|\frac{1}{\eta n}\sum_{i=0}^{n-1} w(i)_j E_2(i)^{(k)}\right| \geq \epsilon\right] \\
&\quad\leq 2\exp\left(-\frac{\epsilon^2(1-\Sigma_{\max}^2)}{2\left(\sqrt{\eta}+1\right)^2}\eta n + \log((s+2r)p+r^2)\right).
\end{aligned}$$

The result follows for $\epsilon = \frac{\theta\lambda_A}{8(4-\theta)}$. This concludes the proof of the lemma. □

Lemma 41. *For*

$$\lambda_A \geq \frac{640(4-\theta)\left\|B^*\right\|_{\infty,1}\left(\frac{\mathcal{D}_{\max}}{\mathcal{C}_{\min}}+1\right)}{\theta D}\sqrt{\frac{\log\left(\frac{4((s+2r)p+r^2)}{\delta}\right)}{n\eta}}$$

with high probability, we have

$$\left\|Y^{(n)}\right\|_\infty \leq \frac{\theta\lambda_A}{4(4-\theta)}.$$

Proof. We can establish

$$Y^{(n)} = \underbrace{B^*\left(R^{(n)}-R^*\right)} + \underbrace{B^*R^*(Q^*)^{-1}\left(Q^*-Q^{(n)}\right)}.$$

We bound these two terms separately. For the first term, we have

$$\left\|B^*\left(R^*-R^{(n)}\right)\right\|_\infty \leq \left\|B^*\right\|_{\infty,1}\left\|\mathcal{Q}^*-\mathcal{Q}^{(n)}\right\|_\infty.$$

For the second term, we have

$$\begin{aligned}
&\left\|B^*R^*(Q^*)^{-1}\left(Q^*-Q^{(n)}\right)\right\|_\infty \\
&\qquad\leq \left\|B^*R^*(Q^*)^{-1}\right\|_{\infty,1}\left\|Q^*-Q^{(n)}\right\|_\infty \\
&\qquad\leq \left\|B^*\right\|_{\infty,1}\sigma_{\max}\left(R^*(Q^*)^{-1}\right)\left\|Q^*-Q^{(n)}\right\|_\infty \\
&\qquad\leq \left\|B^*\right\|_{\infty,1}\frac{\mathcal{D}_{\max}}{\mathcal{C}_{\min}}\left\|Q^*-Q^{(n)}\right\|_\infty.
\end{aligned}$$

The result follows from Lemma 43. This concludes the proof of the lemma.

□

Lemma 42. *For sample complexity*

$$n\eta \geq \frac{3 \times 10^6\, s^3}{D^2\, \theta^2\, \mathcal{C}_{\min}^2} \log\left(\frac{4((s+2r)p + r^2)}{\delta}\right)$$

with high probability, we have

$$\left\|\mathcal{Q}^* - \mathcal{Q}^{(n)}\right\|_\infty \leq \frac{\theta\, \mathcal{C}_{\min}}{9\, s\sqrt{s}}.$$

Proof. Let $X(i) = [x(i)\ u(i)]^T$. Let $\mu(i) = \mathbb{E}\left[X(i)\right]$ (clearly, $\mu(\infty) = 0$). We have

$$\begin{aligned}
&\mathcal{Q}^{(n)} - \mathcal{Q}^* \\
&= \underbrace{\frac{1}{n}\sum_{i=0}^{n-1} \mu(i)\mu(i)^T} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}\left[\left(X(i) - \mu(i)\right)\left(X(i) - \mu(i)\right)^T\right] - \mathcal{Q}^*}_{E_1} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=0}^{n-1} \left(X(i) - \mu(i)\right)\left(X(i) - \mu(i)\right)^T - \left(E_1 + \mathcal{Q}^*\right)}_{E_2}.
\end{aligned}$$

We bound these three terms, separately. For the first term, we have

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=0}^{n-1} \mu(i)\mu(i)^T\right\|_\infty &\leq \frac{1}{n}\sum_{i=0}^{n-1} \Sigma_{\max}^{2i}\left(\|x(0)\|_2^2 + \|u(0)\|_2^2\right) \\
&\leq \frac{\|x(0)\|_2^2 + \|u(0)\|_2^2}{n(1 - \Sigma_{\max}^2)}.
\end{aligned}$$

For the second term, notice that by independency assumption on w, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\left(X(i)-\mu(i)\right)\left(X(i)-\mu(i)\right)^T\right] \\
&\quad=\frac{\eta}{n}\sum_{i=0}^{n-1}\sum_{l=0}^{i-1}\left(I+\eta\mathcal{A}^*\right)^{2l} \\
&\quad=\frac{\eta}{n}\sum_{i=0}^{n-1}\left(I-\left(I+\eta\mathcal{A}^*\right)^{2i}\right)\left(I-\left(I+\eta\mathcal{A}^*\right)^2\right)^{-1} \\
&\quad=\eta\left(\frac{n-1}{n}I-\left(I+\eta\mathcal{A}^*\right)^2+\frac{1}{n}\left(I+\eta\mathcal{A}^*\right)^{2n}\right) \\
&\qquad\qquad\left(I-\left(I+\eta\mathcal{A}^*\right)^2\right)^{-2}.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\mathcal{Q}^* &= \mathbb{E}\left[\lim_{i\to\infty}\left(X(i)-\mu(i)\right)\left(X(i)-\mu(i)\right)^T\right] \\
&= \lim_{i\to\infty}\mathbb{E}\left[\left(X(i)-\mu(i)\right)\left(X(i)-\mu(i)\right)^T\right] \\
&= \lim_{i\to\infty}\eta\sum_{l=0}^{i-1}\left(I+\eta\mathcal{A}^*\right)^{2l} \\
&= \lim_{i\to\infty}\eta\left(I-\left(I+\eta\mathcal{A}^*\right)^{2i}\right)\left(I-\left(I+\eta\mathcal{A}^*\right)^2\right)^{-1} \\
&= \eta\left(I-\left(I+\eta\mathcal{A}^*\right)^2\right)^{-1}.
\end{aligned}$$

In the above inequalities, we interchanged limit and expectation as a result of Gaussianity assumption and the stability of the system. Finally we get

$$\|E_1\|_\infty \le \frac{\eta(1-\Sigma_{\max}^{2n})}{n(1-\Sigma_{\max}^2)^2} \le \frac{\eta}{n(1-\Sigma_{\max}^2)^2}.$$

To bound the third term, notice that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=0}^{n-1}\left(X(i)-\mu(i)\right)\left(X(i)-\mu(i)\right)^T \\
&=\sum_{j=0}^{n-1}\left(I+\eta\mathcal{A}^*\right)^j\left(\frac{n-j}{n}\underbrace{\frac{1}{n-j}\sum_{i=0}^{n-j-1}w(i)w(i)^T}_{V_j}\right)\left(\left(I+\eta\mathcal{A}^*\right)^j\right)^T.
\end{aligned}$$

By Lemma 1 in [113], we have

$$\begin{aligned}&\mathbb{P}\left[\left\|V_j-\eta I\right\|_\infty>\frac{n}{n-j}\epsilon\right]\\&\quad\leq 4\exp\left(-\frac{\epsilon^2 n}{3200\eta(n-j)}n+\log\left((s+2r)p+r^2\right)\right).\end{aligned}$$

Consequently, we get

$$\begin{aligned}&\mathbb{P}\left[\frac{n-j}{n}\Sigma_{\max}^{2(n-j-1)}\left\|V_j-\eta I\right\|_\infty>\Sigma_{\max}^{2(n-j-1)}\epsilon\right]\\&\quad\leq 4\exp\left(-\frac{\epsilon^2 n}{3200\eta(n-j)}n+\log\left((s+2r)p+r^2\right)\right).\end{aligned}$$

Thus, we conclude

$$\begin{aligned}&\mathbb{P}\left[\left\|E_2\right\|_\infty>\frac{1}{1-\Sigma_{\max}^2}\epsilon\right]\\&\qquad\leq 4\exp\left(-\frac{\epsilon^2}{3200\eta}n+\log\left((s+2r)p+r^2\right)\right).\end{aligned}$$

We want this probability to be less than δ. Putting all thre parts together, we get

$$\begin{aligned}&\left\|\mathcal{Q}^*-\mathcal{Q}^{(n)}\right\|_\infty\\&\leq\frac{1}{1-\Sigma_{\max}^2}\left(\frac{\eta(1-\Sigma_{\max}^2)^{-1}+\|x(0)\|_2^2+\|u(0)\|_2^2}{n}+\epsilon\right).\end{aligned}\tag{5.12}$$

For $n\eta\geq\frac{18\,s\sqrt{s}}{D\,\theta\,\mathcal{C}_{\min}}\left(D^{-1}+\|x(0)\|_2^2+\|u(0)\|_2^2\right)$ and $\epsilon=\frac{\eta D\theta\mathcal{C}_{\min}}{18\,s\sqrt{s}}$, the result follows, provide that the probabilities go to zero, i.e.,

$$n\eta\geq\frac{3\times10^6\,s^3}{D^2\,\theta^2\,\mathcal{C}_{\min}^2}\log\left(\frac{4((s+2r)p+r^2)}{\delta}\right).$$

For large enough values of p, this lower bound dominates the earlier lower bound of $n\eta$, hence, we ignore that one. This concludes the proof of the lemma.

□

Lemma 43. *For*

$$\lambda_A \geq \frac{640(4-\theta)\left\|B^*\right\|_{\infty,1}\left(\frac{\mathcal{D}_{\max}}{\mathcal{C}_{\min}}+1\right)}{\theta D}\sqrt{\frac{\log\left(\frac{4((s+2r)p+r^2)}{\delta}\right)}{n\eta}},$$

with high probability, we have

$$\left\|\mathcal{Q}^* - \mathcal{Q}^{(n)}\right\|_\infty \leq \frac{\theta\lambda_A}{4(4-\theta)\ \left\|B^*\right\|_{\infty,1}\left(\frac{\mathcal{D}_{\max}}{\mathcal{C}_{\min}}+1\right)}.$$

Proof. According to (5.12), the result follows if $\epsilon = \frac{\theta\lambda_A D}{8(4-\theta)\|B^*\|_{\infty,1}\left(\frac{\mathcal{D}_{\max}}{\mathcal{C}_{\min}}+1\right)}$ assuming p is large enough.

□

Lemma 44. *For sample complexity*

$$n\eta \geq \frac{3\times 10^6\left(\mathcal{D}_{\max}+2\mathcal{C}_{\min}\right)}{D^2\left(\mathcal{D}_{\max}+\mathcal{C}_{\min}\right)}\log\left(\frac{4((s+2r)p+r^2)}{\delta}\right)$$

with high probability, we have

$$\underbrace{\left\|Q^{(n)}_{\mathcal{S}_k^c\mathcal{S}_k^c} - Q^{(n)}_{\mathcal{S}_k^c\mathcal{S}_k}\left(Q^{(n)}_{\mathcal{S}_k\mathcal{S}_k}\right)^{-1}Q^{(n)}_{\mathcal{S}_k\mathcal{S}_k^c}\right\|_2}_{S^{(n)}} \leq 2(1+\frac{\mathcal{D}_{\max}}{\mathcal{C}_{\min}})\mathcal{D}_{\max}.$$

Proof. Since Q^* and $Q^{(n)}$ are positive semi-definite matrices and

$$\left\|S^{(n)}\right\|_2 \leq \left\|S^{(n)} - S^*\right\|_2 + \left\|S^*\right\|_2$$

The result directly follows from Theorem in [128] for $\epsilon := \|Q^{(n)} - Q^*\|_\infty = \frac{\mathcal{D}_{\max}+\mathcal{C}_{\min}}{4(\mathcal{D}_{\max}+2\mathcal{C}_{\min})}$ considering the fact that $\|S^*\|_2 \leq \mathcal{D}_{\max}\left(1+\frac{\mathcal{D}_{\max}}{\mathcal{C}_{\min}}\right)$.

□

5.11 Proof of the Continuous Time Theorem

Proof. Denote $X(t) = [x(t)\, u(t)]^T$ and let

$$\widehat{\mathcal{Q}} = \frac{1}{T}\int_{t=0}^{T} X(t)X(t)^T dt \qquad \widehat{\mathcal{W}} = \frac{1}{T}\int_{t=0}^{T} dw(t)X(t)^T.$$

Having the result for the discrete time system, it suffices (see proof of Theorem 1.1 in [15] for more details) to show that for a given continuous time system, there exists a discrete time system with $\mathcal{Q}^{(n)}$ and $\mathcal{W}^{(n)}$ such that almost surely,

$$\mathcal{Q}^{(n)} \longrightarrow \widehat{\mathcal{Q}} \qquad\qquad \mathcal{W}^{(n)} \longrightarrow \widehat{\mathcal{W}},$$

as $n \to \infty$ for a fixed $T = n\eta$ (and hence, $\eta \to 0$).

Let $\mathcal{Q}^*$ be the matrix satisfying the continuous time Lyapunov stability equation $\mathcal{A}^*\mathcal{Q}^* + \mathcal{Q}^*\mathcal{A}^{*T} + I = 0$ and $\mathcal{Q}^*(\eta)$ be the matrix satisfying the discrete time Lyapunov stability equation $\mathcal{A}^*\mathcal{Q}^*(\eta) + \mathcal{Q}^*(\eta)\mathcal{A}^{*T} + \eta\mathcal{A}^*\mathcal{Q}^*(\eta)\mathcal{A}^{*T} + I = 0$. It is easy to see that $\mathcal{Q}^*(\eta) \to \mathcal{Q}^*$ as $\eta \to 0$ by the uniqueness of the stationary distribution. Moreover, by Lemma 42, we know that $\mathcal{Q}^{(n)} \to \mathcal{Q}^*(\eta)$ as $n \to \infty$.

Now, let the initial state of the discrete time system be

$$X(i=0) = \left(\mathcal{Q}^*(\eta)\right)^{1/2} \left(\mathcal{Q}^*\right)^{-1/2} X(t=0),$$

and the noise $w(i) = w(t = i\eta) - w(t = (i-1)\eta)$. It can be easily checked that $w(i) \sim \mathcal{N}(0, \eta I)$ if the continuous time $w(t)$ is a Brownian motion. Thus, $x(i)$ and $x(t)$ are coupled and the almost sure convergence, follows from the convergence of random walks to Brownian motions [91]. This concludes the proof of the theorem for continuous time systems.

□

Chapter 6

Greedy Dirty Models

This chapter considers the multiple sparse linear regression problem (multi-task problem) in high-dimensional setting (from small number of observations). It has recently been shown that by taking advantage of partially shared supports across tasks, it is possible to lower the sample complexity for sparsity structure recovery by using a "dirty model": a super-position of sparse and group-sparse modeling approaches; this is based on convex regularization. In this chapter, we provide a new forward-backward greedy procedure. Each forward step involves the addition of either a shared feature common to all tasks, or a unique feature to one of the tasks, chosen in a natural greedy fashion. Each backward step involves greedy removal of potentially multiple features of either type. We provide a statistical guarantee for the performance of the greedy algorithm which is identical to convex optimization guarantees with significantly milder assumptions. Empirical evidence on synthetic and real data shows that our algorithm outperforms all convex approaches, in terms of both sample complexity (requiring fewer samples for structure recovery), and computational complexity.

6.1 Introduction

Multi-task Learning: In many applications such as clustering, classification, regression, etc, we observe some data over a superset of features and we want to learn different concepts (tasks), each depending on a subset of features. One question to ask here is that if there is an advantage in learning all concepts *jointly* as a single problem as opposed to learning them as separate problems. Learning the relevant subsets of features for all tasks jointly is often called *multi-task learning* [26]. We consider the problem of multi-task learning when our observations are noisy linear measurements known as *multiple*

linear regression. Multiple linear regression comes up in different applications ranging from graphical model selection [112], kernel learning [7], function estimation [110], etc. The setup is as follows: There are r target tasks (sparse vectors in this case) $\beta^{*(1)}, \ldots, \beta^{*(r)} \in \mathbb{R}^p$ to learn and for each task j, we observe n_j noisy linear measurements according to the statistical model

$$y^{(j)} = X^{(j)}\beta^{*(j)} + w^{(j)} \qquad \forall j \in \{1, \ldots, r\}, \tag{6.1}$$

where, $X^{(j)} \in \mathbb{R}^{n_i \times p}$ is the j^{th} design matrix, $y^{(j)} \in \mathbb{R}^{n_j}$ is the j^{th} response vector and $w^{(j)} \in \mathbb{R}^{n_j}$ is the iid zero mean Gaussian noise with variance σ^2. We combine all tasks $\beta^{*(j)}$ as columns of a matrix $\beta^* \in \mathbb{R}^{p \times r}$. The problem here is that given $y^{(j)}$'s and $X^{(j)}$'s, estimate the matrix β^*.

High-dimensional Setting: With expensive or lengthy measurement processes, we increasingly face situations where the number of observations n_j is substantially smaller than the number of features p; normally $n_j \propto \log(p)$. Since the problem is ill-posed under this setting, we often impose some extra structure on the target parameters. Some of the popular structures include sparsity/block sparsity (compressed sensing [13], LASSO [131], group sparsity [151]), low-rank [102, 116], sparse/block sparse Markov random fields [70, 112], Hankel structure [49], etc. In all these cases, we assume that the target parameter lies in a low-dimensional subspace of size $\log(p)$ and hence $\mathcal{O}(\log(p))$ observations should suffice for finding the target parameter. In our setup, we assume the matrix β^* is a sparse matrix, i.e., each task depends only on a small number of features.

Convex Optimization Methods: There are three main approaches to solve the multiple sparse regression problems: a) Considering each task separately and using the ℓ_1-norm on each task to impose sparsity structure known as LASSO [142]; b) Focusing only on shared features (rows of β^* that are mostly non-zero element-wise) and using group sparsity regularizer such as ℓ_1/ℓ_∞-norm [101, 136] or ℓ_1/ℓ_2-norm [89, 104] to impose block sparsity structure; c) Considering the matrix β^* as a superposition of a block sparse matrix B^* containing rows corresponding to shared features and a sparse matrix S^* containing rows corresponding to non-shared features and use group sparse

regularizer on B^* and sparse regularizer on S^* known as "dirty model" [69] method. Depending on the number of shared features, each of the first and second approaches have advantage over each other; but dirty model outperforms both approaches under all sharedness regimes.

In all these methods, the convex regularizer is a surrogate convex proxy to the structure. For example, ℓ_1-norm is a convex surrogate for the number of non-zero elements known as ℓ_0-norm[1]. Because of this convexification, any structure recovery guarantee using these methods require strong assumptions on the Hessian of the loss function, e.g. irrepresentable condition, to ensure that the surrogate convex regularizer does not bias the loss function. These assumptions are necessary [142] and known to be hard to satisfy in practice.

Although convex optimization problems can be efficiently solved in polynomial time, still their computational complexity is high. As an example, the computational cost for a convex optimization problem is typically $\mathcal{O}(p^4)$ which is intractable for a typical data mining application with $p = 10^6$.

Greedy Methods: In contrast with convex optimization methods, there are has been some greedy proposals based on Orthogonal Matching Pursuit (aka greedy least square regression, forward greedy selection) [135, 154]. These forward greedy algorithms select the next best feature and add it to the set of active features and then optimize the loss function over the set of active features. Typically, the search for the next best feature can be done in parallel and also optimization step is only over a small set of active features which lead to significant speedup. It has been shown that these algorithms, although are faster, require similar assumptions to the convex optimization methods to succeed.

Recently, [152] introduced a forward-backward greedy algorithm to find the sparse solution of the squared loss. They provide the same statistical guarantee as the convex optimization without need to strong assumptions. With slightly different analysis, [68] showed the same guarantee for any general loss. These algorithms seem to be very promising for simple models, but they can-

[1]This is not a norm since it does not satisfy the triangle inequality.

not handle dirty models.

Our Contribution: We provide a novel forward-backward greedy algorithm for dirty models, i.e., when the target structure is a superposition of a sparse and block-sparse matrix. We provide theoretical guarantee on the performance of the algorithm in terms of both estimation error and support recovery. Our analysis is more subtle comparing to [68], since we would like to have *local* assumptions on each task $\beta^{*(j)}$ as opposed to having *global* assumptions on the whole matrix β^*.

Dealing with group sparsity, one of the issues with convex optimization methods is that it is not clear which ℓ_1/ℓ_p-norm should be chosen as regularizer. The analysis for $p = \infty$ [101] and $p = 2$ [104] show that the sample complexity required for structure recovery depends on the choice of the regularizer. These greedy approaches solve this problem by dealing directly with features as opposed to dealing with their relaxations and empirical results show significant improvements in sample complexity.

6.2 Greedy Algorithm for Dirty Model

Considering the loss function

$$\mathcal{L}(\beta) = \sum_{j=1}^{r} \frac{1}{2n_j} \|y^{(j)} - X^{(i)}\beta^{(j)}\|_2^2,$$

Algorithm 6 is forward-backward greedy proposal for dirty models. This algorithm inputs data, two stopping parameters ϵ_s, ϵ_b and a backward factor ν and outputs the estimate $\widehat{\beta}$ which is a superposition of a sparse matrix supported on $\widehat{\mathbb{S}}_s$ and a block-sparse matrix supported on $\widehat{\mathbb{S}}_b$. In the next section, we show that ϵ_s, ϵ_b must be chosen such that $1 \leq \frac{\epsilon_b}{\epsilon_s} \leq r$. If the loss function is curved in many directions around the optimal point, ν can be chosen closer to zero to get speedup, but practically, it seems that $\nu = \frac{1}{2}$ is a good choice [152]. We quantify this curvature in the next section.

Identifiability: The matrix β^* can be written as superpositions of many pairs of sparse and block-sparse matrices and hence, it is not clear what

Algorithm 6 Greedy forward-backward algorithm for finding a sparse + block-sparse optimizer of $\mathcal{L}(\cdot)$

Input: Data $D := \{y^{(1)}, X^{(1)}, \ldots, y^{(r)}, X^{(r)}\}$, Stopping Thresholds ϵ_b and ϵ_s, Backward Factor $\nu \in (0,1)$

Output: Sparse + Block-sparse Optimizer $\widehat{\beta}$

$\widehat{\beta}(0) \longleftarrow \mathbf{0}$ and $\underline{\widehat{\mathbb{S}}}_b^{(0)}, \widehat{\mathbb{S}}_b^{(0)}, \widehat{\mathbb{S}}_s^{(0)} \longleftarrow \phi$ and $\mu_b^{(0)}, \mu_s^{(0)} \longleftarrow 0$ and $k_b, k_s, k \longleftarrow 1$

while true **do** {*Forward Step*}

$(i_*^b, \alpha_*) \longleftarrow \underset{i \notin \underline{\widehat{\mathbb{S}}}_b^{(k_b-1)}; \alpha \in \mathbb{R}^r}{\arg\min} \mathcal{L}(\widehat{\beta}(k-1) + e_i \alpha^T; D)$ and $\mu_b^{(k)} \longleftarrow \frac{\mathcal{L}(\widehat{\beta}(k-1);D) - \mathcal{L}\left(\widehat{\beta}(k-1) + e_{i_*^b} \alpha_*^T; D\right)}{\epsilon_b}$

$((i_*^s, j_*^s), \gamma_*) \longleftarrow \underset{(i,j) \notin \widehat{\mathbb{S}}_s^{(k-1)}; \gamma \in \mathbb{R}}{\arg\min} \mathcal{L}(\widehat{\beta}(k-1) + \gamma e_i e_j^T; D)$ and $\mu_s^{(k)} \longleftarrow \frac{\mathcal{L}(\widehat{\beta}(k-1);D) - \mathcal{L}\left(\widehat{\beta}(k-1) + \gamma_* e_{i_*^s} e_{j_*^s}^T; D\right)}{\epsilon_s}$

if $\max(\mu_s^{(k)}, \mu_b^{(k)}) \leq 1$ **then**
  break
end if

if $\mu_b^{(k)} \geq \mu_s^{(k)}$ **then**
  $\underline{\widehat{\mathbb{S}}}_b^{(k_b)} \longleftarrow \underline{\widehat{\mathbb{S}}}_b^{(k_b-1)} \cup \{i_*^b\}$ and $\widehat{\mathbb{S}}_b^{(k_b)} \longleftarrow \widehat{\mathbb{S}}_b^{(k_b-1)} \cup \{(i_*^b, j) : \forall j\}$ and $k_b \longleftarrow k_b + 1$
else
  $\widehat{\mathbb{S}}_s^{(k_s)} \longleftarrow \widehat{\mathbb{S}}_s^{(k_s-1)} \cup \{(i_*^s, j_*^s)\}$ and $k_s \longleftarrow k_s + 1$
end if

$\widehat{\beta}(k) \longleftarrow \arg\min_{\beta} \mathcal{L}\left(\beta_{\widehat{\mathbb{S}}_b^{(k_b-1)} \cup \widehat{\mathbb{S}}_s^{(k_s-1)}}; D\right)$
$k \longleftarrow k + 1$

while true **do** {*Backward Step*}

$\mu^{(k-1)} \longleftarrow \max\left(\mu_b^{(k-1)}, \mu_s^{(k-1)}\right)$
$i_b^* \longleftarrow \underset{i \in \underline{\widehat{\mathbb{S}}}_b^{(k_b-1)}}{\arg\min} \mathcal{L}(\widehat{\beta}(k-1) - e_i \widehat{\beta}_i(k-1); D)$ and $\nu_b \longleftarrow \frac{\mathcal{L}\left(\widehat{\beta}(k-1) - e_{i_b^*} \widehat{\beta}_{i_b^*}(k-1); D\right) - \mathcal{L}(\widehat{\beta}(k-1); D)}{\mu^{(k-1)} \epsilon_b}$
$(i_s^*, j_s^*) \longleftarrow \underset{(i,j) \in \widehat{\mathbb{S}}_s^{(k_s-1)}}{\arg\min} \mathcal{L}(\widehat{\beta}(k-1) - \widehat{\beta}_i^{(j)}(k-1) e_i e_j^T; D)$ and $\nu_s \longleftarrow \frac{\mathcal{L}\left(\widehat{\beta}(k-1) - \widehat{\beta}_{i_s^*}^{(j_s^*)}(k-1) e_{i_s^*} e_{j_s^*}^T; D\right) - \mathcal{L}(\widehat{\beta}(k-1); D)}{\mu^{(k-1)} \epsilon_s}$

sparse matrix and what block-sparse matrix is our target in the algorithm. Here, we discuss that the choice of the target sparse and block-sparse matrix depends on the ratio $d = \lceil \frac{\epsilon_b}{\epsilon_s} \rceil$. Split $\beta^* = B^* + S^*$, where, B^* includes the rows with more than or equal to d non-zeros and S^* includes the rows with strictly less than d non-zeros. Let $\mathcal{S}_s^*$ to be the support of S^* and $\underline{\mathcal{S}}_b^*$ to be the row-support of B^*. Our algorithms tries to find $\widehat{\mathcal{S}}_s \approx \widehat{S}_s^*$ and $\widehat{\underline{\mathcal{S}}}_b \approx \widehat{\underline{S}}_b^*$. Notice that we are not looking for any specific sparse or block-sparse matrix, rather, we are looking for their superposition. Thus, we can search for different values of ϵ_b, ϵ_s and as long as $\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_b = \widehat{S}_s^* \cup \widehat{S}_b^*$, our support recovery is successful. Here, $\widehat{\mathcal{S}}_b = \{(i,j) : i \in \widehat{\underline{\mathcal{S}}}_b\}$ and $\mathcal{S}_b^* = \{(i,j) : i \in \underline{\mathcal{S}}_b^*\}$.

Forward Stop: The algorithm starts with an empty set of active blocks (rows) $\widehat{\underline{S}}_b^{(0)}$ and single elements $\widehat{S}_s^{(0)}$. At each forward step, the algorithm finds the next best row and the next best single element as potential candidates to be added to the active sets. Then, we normalize the loss function improvement of each candidate over ϵ_b, ϵ_s and add the best to the active set. We conclude the forward step by re-optimizing the parameter over the union of the updated set of active rows/elements $\widehat{\mathcal{S}}_s^{(k)} \cup \widehat{\mathcal{S}}_b^{(k)}$.

Backward Stop: In the backward step, the algorithm finds the worst (in terms of loss function improvement) row and worst single element. Similar to the forward step, we normalize the improvement over ϵ_b, ϵ_s and remove the worst of the two. We keep repeating the backward step, until the active sets contain rows/single elements that each of then provide significant improvement to the loss function.

Convergence: The parameter ν ensures that at each iteration consisting of a single forward step and potentially multiple backward steps, the loss function is improved by at least $(1-\nu)\epsilon_s$ (if a single element is selected in the forward step) and $(1-\nu)\epsilon_b$ (if a row is selected in the forward step). This shows that the algorithm stops after finite iterations.

We analyze the algorithm and provide theoretical guarantee on its performance in the next section.

6.3 Theoretical Guarantee

The goal of our theoretical analysis is to provide an estimation error bound as well as a sparsistency guarantee by imposing some assumptions on the loss function. We only require that the loss function have some curvature around the optimal point and formalize this assumption as restricted eigenvalue property.

Restricted Eigenvalue Property: We say the Hessian matrix $Q^{(j)} = X^{(j)}X^{(j)T}$ satisfies $REP(s_j)$ if the restricted eigenvalue property (REP) on s_j-sparse vectors $\delta \in \mathbb{R}^p$ holds with constants $C_{\min}$ and $\rho \geq 1$; that is

$$C_{\min}\|\delta\|_2 \leq \|Q^{(j)}\delta\|_2 \leq \rho C_{\min}\|\delta\|_2 \qquad \forall \|\delta\|_0 \leq s \tag{6.2}$$

Without loss of generality, we assume that $C_{\min}$ and ρ are the same for all tasks. We can provide a guarantee for the case that the design matrix $X^{(j)}$ is Gaussian and the population Hessian matrix satisfies REP.

Lemma 45. *If each row of the design matrix $X^{(j)} \in \mathbb{R}^{n\times p}$ is distributed as $\mathcal{N}(0, \Sigma^{(j)})$ and $\Sigma^{(j)}$ satisfies $REP(s_j)$, then for any small $\theta > 0$, the matrix $Q^{(j)} = X^{(j)}X^{(j)T}$ satisfies*

$$(1-\theta)C_{\min}\|\delta\|_2 \leq \|Q^{(j)}\delta\|_2 \leq (1+\theta)\rho C_{\min}\|\delta\|_2, \tag{6.3}$$

for all $\|\delta\|_0 \leq s_j$, with probability $1-c_1\exp(-c_2 n)$ provided that $n_j \geq c_3(\theta)\, s_j \log(p)$, where c_1, c_2 and c_3 are constants independent of (n_j, s_j, p).

The proof follows from Lemma 9 (Appendix K) in [142]. This lemma shows that for Gaussian design matrices, $REP(s_j)$ is satisfied with high probability for $\mathcal{O}(s_j \log(p))$ samples.

Gradient of the loss function: Since β^* is the asymptotic optimal point of the loss function, we have $\mathbb{E}[\nabla\mathcal{L}(\beta^*)] = 0$. With finite number of observations, the gradient $\nabla^{(j)} = -X^{(j)T}\left(y^{(j)} - X^{(j)}\beta^{*(j)}\right)$ will not be zero at β^* due to the noise. However, we can bound this quantity in the following lemma if the number of samples is not too small.

Lemma 46. *Given the sample complexity $n_j \geq c_4 \log(rp)$ for some constant c_4 and all $j \in \{1, 2, \ldots, r\}$, we have*

$$\max_j \|\nabla^{(j)}\|_\infty \leq c_5 \sqrt{\frac{\log(rp)}{n}} := \lambda_n$$

with probability at least $1 - c_6 \exp(-c_7 n)$ for some constants c_5, c_6 and c_7 independent of (n_j, s_j, p).

The proof follows from Lemma 5 in [142]. We state our theoretical result in terms of λ_n for the sake of generality. This parameter can be replaced with any upper-bound on $\nabla^{(j)}$ and our guarantee still holds.

Finally, we define $s_j^* = |\{(i,j) : (i,j) \in \mathcal{S}_s^* \cup \mathcal{S}_b^*\}|$ to be the size of the support of the j^{th} task. Now, we can state our guarantee on the performance of the Algorithm 6.

Theorem 10 (Sparsistency)**.** *Suppose $1 \leq \frac{\epsilon_b}{\epsilon_s} \leq r$ and $REP\left(\eta\, s_j^*\right)$ holds for some $\eta \geq 2 + \frac{4r\epsilon_s\rho^4(\rho^4-\rho^2+2)}{\epsilon_b\nu}$. Provided the sample complexity $n_j \geq K s_j^* \log(rp)$ for some constant K, if we run Algorithm 6 with stopping threshold $\epsilon_s \geq \frac{4\rho^2\eta r s^* \lambda_n^2}{\nu C_{\min}^2}$ and β^* satisfies $\min_{(i,j)\in\mathcal{S}_s^*\cup\mathcal{S}_b^*} |\beta_i^{*(j)}| > \frac{4\rho}{C_{\min}}\sqrt{\epsilon_s}$, the output $\widehat{\beta}$ with support $\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s$ satisfies:*

(a) **Error Bound:** $\|\widehat{\beta} - \beta^*\|_F \leq \frac{\sqrt{rs^*}}{C_{\min}}\left(\frac{\lambda_n\sqrt{\eta}}{C_{\min}} + 2\rho\sqrt{\epsilon_s}\right)$.

(b) **No False Exclusions:** $\mathcal{S}_b^* \cup \mathcal{S}_s^* \;-\; \widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s = \emptyset$.

(c) **No False Inclusions:** $\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s \;-\; \mathcal{S}_b^* \cup \mathcal{S}_s^* = \emptyset$.

The result holds with high-probability for Gaussian designs when the population matrices $\Sigma^{(i)}$ satisfy REP.

Remark 1. This theorem provides stronger results than [68]. Here, our assumption is local – that is we assume REP for each individual task as opposed to a global assumption that assumes REP for the whole set of tasks. The global assumption requires $REP(\eta \sum s_j)$ for each task, which is orderwise (by an order of r) worse than our assumption $REP(\eta s_j)$ for task j. To get this tight result, we have a careful per-task analysis detailed in the appendix.

Proof. The proof technique is inspired by [68]. Provided lemmas 47, 48 and 49 hold, we show below that the greedy algorithm is sparsistent. However, these lemmas require *apriori* that the REP condition hold for sparsity size $\widehat{s}_j$ where $\widehat{s}_j$ is the support size of the j^{th} columns among all matrices supported on $\mathcal{S}_b^* \cup \mathcal{S}_s^* \cup \widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s$. Thus, we use the result in Lemma 50 that if $RSC(\eta s_j^*)$ holds, then the solution when the algorithm terminates satisfies $s_j \leq (\eta - 1)s_j^*$, where, s_j is the maximum support size of the columns of $\widehat{\beta}$ and hence $\widehat{s}_j \leq \eta s_j^*$. Thus, we can then apply Lemmas 47, 49 and Lemma 48 to complete the proof as detailed below.

(a) The result follows directly from Lemma 48, and noting that $s \leq \eta s^*$.

(b) We follow the chaining argument in [152]. For any $\tau \in \mathbb{R}$, we have

$$\begin{aligned}\tau \left| \{(i,j) \in (\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s) : |\beta_i^{*(j)}|^2 > \tau\} \right| &\leq \|\beta^*_{(\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s)}\|_F^2 \leq \|\beta^* - \widehat{\beta}\|_F^2 \\ &\leq \frac{2\eta r s^* \lambda_n^2}{C_{\min}^4} + \frac{8\rho^2 \epsilon_s}{C_{\min}^2} |(\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s)|, \end{aligned} \tag{6.4}$$

where the last inequality follows from part (a) and the inequality $(a+b)^2 \leq 2a^2 + 2b^2$. Now, setting $\tau = \frac{16\rho^2\epsilon_s}{C_{\min}^2}$, and dividing both sides by $\tau/2$ we get

$$2 \left| \{(i,j) \in (\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s) : |\beta_i^{*(j)}|^2 > \tau\} \right| \leq \frac{\eta r s^* \lambda_n^2}{4\rho^2 C_{\min}^2 \epsilon_s} + |(\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s)|. \tag{6.5}$$

Substituting $\left| \{(i,j) \in (\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s) : |\beta_i^{*(j)}|^2 > \tau\} \right| = |(\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s)| - \left| \{(i,j) \in (\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s) : |\beta_i^{*(j)}|^2 \leq \tau\} \right|$, we get

$$\begin{aligned}|(\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s)| &\leq 2 \left| \{(i,j) \in (\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s) : |\beta_i^{*(j)}|^2 \leq \tau\} \right| + \frac{\eta r s^* \lambda_n^2}{4\rho^2 C_{\min}^2 \epsilon_s} \\ &\leq 2 \left| \{(i,j) \in (\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s) : |\beta_i^{*(j)}|^2 \leq \tau\} \right| + 1/2, \end{aligned}$$

due to the setting of the stopping threshold ϵ_s. This in turn entails that

$$|(\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s)| \leq 2 \left| \{(i,j) \in (\mathcal{S}_b^* \cup \mathcal{S}_s^*) - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s) : |\beta_i^{*(j)}|^2 \leq \tau\} \right| = 0, \tag{6.6}$$

by our assumption on the size of the minimum entry of β^*.

(c) From Lemma 49 and the result of Part (b), we have

$$|(\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s) - (\mathcal{S}_b^* \cup \mathcal{S}_s^*)| \leq \frac{\rho^2 C_{\min}^2}{\nu \epsilon_s} \left\| \widehat{\beta}_{(\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s) - (\mathcal{S}_b^* \cup \mathcal{S}_s^*)} \right\|_F^2 \leq \frac{\rho^2 C_{\min}^2}{\nu \epsilon_s} \left\| \widehat{\beta} - \beta^* \right\|_F^2$$
$$\leq \frac{\rho^2 C_{\min}^2}{\nu \epsilon_s} \frac{2\eta r s^* \lambda_n^2}{C_{\min}^4} \leq 1/2$$

due to the setting of the stopping threshold ϵ_s. □

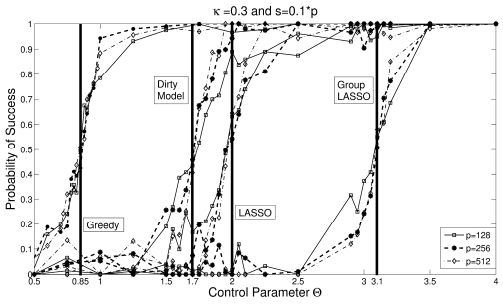
6.4 Experimental Results

In this section we compare our greedy algorithm with other convex-optimization based methods for multi-task learning. In particular, we compare our algorithm against LASSO, group LASSO (with ℓ_1/ℓ_∞ regularizer) and dirty model [69]. We get substantially better results both on synthetic and real data.

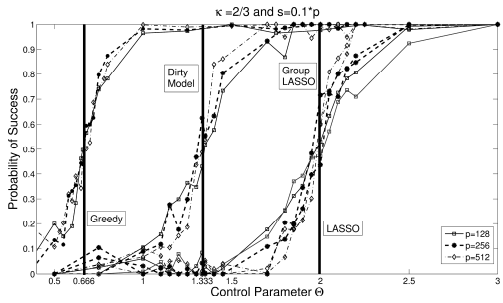
6.4.1 Synthetic Data

To have a common ground for comparison, we run the same experiment used for the comparison of LASSO, group LASSO and dirty model in [69, 101]. Consider the case where we have $r = 2$ tasks each with the support size of $s = p/10$ and suppose these two tasks share a κ portion of their supports. The location of non-zero entries are chosen uniformly at random and values of β_1^* and β_2^* are chosen to be standard Gaussian realizations. Each row of he matrices $X^{(1)}$ and $X^{(2)}$ is distributed as $\mathcal{N}(0, I)$ and each entry of the noise vectors w_1 and w_2 is a zero-mean Gaussian draw with variance 0.1. We run the experiment for problem sizes $p \in 128, 256, 512$ and for support overlap levels $\kappa \in 0.3, 2/3, 0.8$.

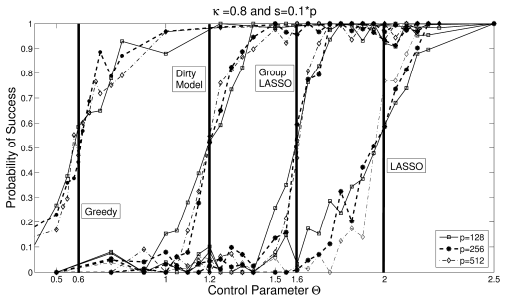
We use cross-validation to find the best values of regularizer coefficients. To do so, we choose $\epsilon_s = c\,\frac{s\log(p)}{n}$, where $c \in [10^{-4}, 10]$, and $\epsilon_b = k\,\epsilon_s$, where $k \in [1, 2]$. Notice that this search region is motivated by the requirements of our theorem and can be substantially smaller than the region needs to be searched for ϵ_s and ϵ_b if they are independent. Interestingly, for small number of samples n, the ratio k tends to be close to 1, where for large number of



(a) Little support overlap: $\kappa = 0.3$



(b) Moderate support overlap: $\kappa = 2/3$



(c) High support overlap: $\kappa = 0.8$

Figure 6.1: Probability of success in recovering the exact sign support using greedy algorithm, dirty model, Lasso and group LASSO (ℓ_1/ℓ_∞). For a 2-task problem, the probability of success for different values of feature-overlap fraction κ is plotted. Here, we let $s = p/10$ and the values of the parameter and design matrices are i.i.d standard Gaussians. Also, the noise variance is set to be $\sigma = 0.1$. As we can see, greedy method outperforms all methods in the minimum number of samples required for sign support recovery.

samples, the ratio tends to be close to 2. We suspect this phenomenon is due to the lack of curvature around the optimal point when we have few samples. The greedy algorithm is more stable if it picks a row as opposed to a single coordinate, even if the improvement of the entire row is comparable to the improvement of a single coordinate.

To compare different methods under this regime, we define a re-scaled version of sample size n, aka control parameter Θ, as follows:

$$\Theta = \frac{n}{s \log (p - (2 - \kappa)s)}.$$

For different values of κ, we plot the probability of success, obtained by averaging over 100 problems, versus the control parameter Θ in Fig.6.4. It can be seen that the greedy method outperforms, i.e., requires less number of samples, to recover the exact sign support of β^*.

This result matches the known theoretical guarantees. It is well-known that LASSO has a sharp transition at $\Theta \approx 2$ [142][1], group LASSO (ℓ_1/ℓ_∞ regularizer) has a sharp transition at $\Theta = 4 - 3\kappa$ [101] and dirty model has a sharp transition at $\Theta = 2 - \kappa$ [69]. Although we do not have a theoretical result, these experiments suggest the following conjecture:

Conjecture 1. *For two-task problem with $C_{\min} = \rho = 1$ and Gaussian designs, the greedy algorithm has a sharp transition at $\Theta = 1 - \frac{\kappa}{2}$.*

To investigate our conjecture, we plot the sharp transition thresholds for different methods versus different values of $\kappa \in \{0.05, 0.3, 2/3, 0.8, 0.95\}$ for problem sizes $p \in \{128, 256, 512\}$. Fig 6.2 shows that the sharp transition threshold for greedy algorithm follows our conjecture with a good precision. Although, theoretical guarantee for such a tight threshold remains open.

6.4.2 Handwritten Digits Dataset

We use the handwritten digit dataset [44] that is used by a number of papers [60, 69, 108] as a reliable dataset for optical handwritten digit recogni-

[1]The exact expression is $\frac{n}{s \log(p)} = 2$. Here, we ignore the term $(2 - \kappa)s$ comparing to p.

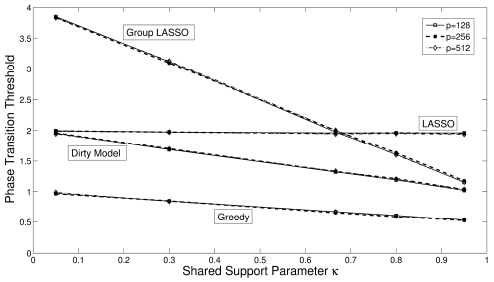


Figure 6.2: Behavior of phase transition threshold versus the parameter κ in a 2-task problem for greedy algorithm, dirty model, LASSO and group LASSO (ℓ_1/ℓ_∞ regularizer). The y-axis is $\Theta = \frac{n}{s\log(p-(2-\kappa)s)}$, where n is the number of samples at which threshold was observed. Here, we let $s = p/10$ and the values of the parameter and design matrices are i.i.d standard Gaussians. Also, the noise variance is set to be $\sigma = 0.1$. The greedy algorithm shows substantial improvement in terms of the minimum number of samples required for exact sign support recovery over the other methods.

tion algorithms. The dataset contains $p = 649$ features of handwritten numerals 0-9 ($r = 10$ tasks) extracted from a collection of Dutch utility maps. The dataset provides 200 samples of each digit written by different people. We take $n/10$ samples from each digit and combine them to a big matrix $X \in \mathbb{R}^{n\times p}$, i.e., we set $X^{(i)} = X$ for all $i \in \{1, \ldots, 10\}$. We construct the response vectors y_i to be 1 if the corresponding row in X is an instance of i^{th} digit and zero otherwise. Clearly, y_i's will have a disjoint support sets. We run all four algorithms on this data and report the results.

Table 6.1 shows the results of our analysis for different sizes of the training set n. We measure the classification error for each digit to get the 10-vector of errors. Then, we find the average error and the variance of the error vector to show how the error is distributed over all tasks. Again, in all methods, parameters are chosen via cross-validation. It can be seen that the greedy method provides a more consistent model selection as the model complexity does not change too much as the number of samples increases while the classification error decreases substantially. In all cases, we get $\%25 - \%30$ improvement in classification error.

n		Greedy	Dirty Model	Group LASSO	LASSO
10	Average Classification Error	6.5%	8.6%	9.9%	10.8%
	Variance of Error	0.4%	0.53%	0.64%	0.51%
	Average Row Support Size	180	171	170	123
	Average Support Size	1072	1651	1700	539
20	Average Classification Error	2.1%	3.0%	3.5%	4.1%
	Variance of Error	0.44%	0.56%	0.62%	0.68%
	Average Row Support Size	185	226	217	173
	Average Support Size	1120	2118	2165	821
40	Average Classification Error	1.4%	2.2%	3.2%	2.8%
	Variance of Error	0.48%	0.57%	0.68%	0.85%
	Average Row Support Size	194	299	368	354
	Average Support Size	1432	2761	3669	2053

Table 6.1: Handwriting Classification Results for greedy algorithm, dirty model, group LASSO and LASSO. The greedy method provides much better classification errors with simpler models. The greedy model selection is more consistent as the number of samples increases.

6.5 Auxiliary Lemmas for Theorem 10

In this section, we prove the Lemmas used in the proof of Theorem 1. We use $\widehat{\underline{\mathbb{S}}}_b$, $\widehat{\underline{\mathbb{S}}}_s$, $\underline{\mathbb{S}}_b^*$ and $\underline{\mathbb{S}}_s^*$ to represent the row-supports (indecies of the rows with at least one non-zero entry). Let

$$M^* = \{(i,j) : j \in \underline{\mathbb{S}}_b^* \cap \widehat{\underline{\mathbb{S}}}_s;\ (i,j) \notin \widehat{\mathbb{S}}_s\}$$
$$\widehat{M} = \{(i,j) : j \in \widehat{\underline{\mathbb{S}}}_b \cap \underline{\mathbb{S}}_s^*;\ (i,j) \notin \mathbb{S}_s^*\}$$

It is clear that

$$\begin{aligned} |\widehat{\underline{\mathbb{S}}}_b - (\underline{\mathbb{S}}_b^* \cup \underline{\mathbb{S}}_s^*)|\epsilon_b &\leq \left(|\widehat{\mathbb{S}}_b - (\mathbb{S}_b^* \cup \mathbb{S}_s^*)| - |\widehat{M}|\right)\epsilon_s \\ |\underline{\mathbb{S}}_b^* - (\widehat{\underline{\mathbb{S}}}_b \cup \widehat{\underline{\mathbb{S}}}_s)|\epsilon_b &\leq \left(|\mathbb{S}_b^* - (\widehat{\mathbb{S}}_b \cup \widehat{\mathbb{S}}_s)| - |M^*|\right)\epsilon_s \end{aligned},$$

since $\frac{\epsilon_b}{r} \leq \epsilon_s$.

Note that when the algorithm terminates, the forward step fails to go through. This entails that

$$\begin{aligned} \mathcal{L}(\widehat{\beta}) \quad - \inf_{i \notin \widehat{\underline{\mathbb{S}}}_b, \alpha \in \mathbb{R}^r} \mathcal{L}(\widehat{\beta} + e_i\, \alpha^T) < \epsilon_b \\ \mathcal{L}(\widehat{\beta}) - \inf_{(i,j) \notin \widehat{\mathbb{S}}_s, \gamma \in \mathbb{R}} \mathcal{L}(\widehat{\beta} + \gamma e_i e_j^T) < \epsilon_s. \end{aligned}$$

The next lemma shows that this has the consequence of upper bounding the deviation in loss between the estimated parameters $\widehat{\beta}$ and the true parameters β^*.

Lemma 47 (Stopping Forward Step)**.** *When the algorithm stops with parameter $\widehat{\beta}$, we have*

$$\left|\mathcal{L}\left(\widehat{\beta}\right) - \mathcal{L}\left(\beta^*\right)\right| < 2\rho C_{\min}\sqrt{|(\mathcal{S}_s^* \cup \mathcal{S}_b^*) - (\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_s)|\epsilon_s}\ \left\|\widehat{\beta} - \beta^*\right\|_F. \tag{6.7}$$

Proof. Let $\widehat{\Delta} = \beta^* - \widehat{\beta}$. For any $\eta \in \mathbb{R}$, we have

$$\begin{aligned}
&- |(\mathcal{S}_s^* \cup \mathcal{S}_b^*) - (\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_s)|\epsilon_s \\
&\qquad = -|\mathcal{S}_s^* - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s)|\epsilon_s - |\mathcal{S}_b^* - (\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_s)|\epsilon_s \\
&\qquad \leq -|\mathcal{S}_s^* - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s)|\epsilon_s - |M^*|\epsilon_s - |\underline{\mathcal{S}}_b^* - (\widehat{\underline{\mathcal{S}}}_b \cup \widehat{\underline{\mathcal{S}}}_s)|\epsilon_b \\
&\qquad < \sum_{(i,j)\in(\mathcal{S}_s^* - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s))\cup M^*} \left(\mathcal{L}\left(\widehat{\beta} + \eta\widehat{\Delta}_i^{(j)} e_i e_j^T\right) - \mathcal{L}\left(\widehat{\beta}\right)\right) + \sum_{i\in\underline{\mathcal{S}}_b^* - (\widehat{\underline{\mathcal{S}}}_b \cup \widehat{\underline{\mathcal{S}}}_s)} \left(\mathcal{L}\left(\widehat{\beta} + \eta e_i\widehat{\Delta}_i\right) - \mathcal{L}\left(\widehat{\beta}\right)\right) \\
&\qquad \leq \eta\left\langle \nabla\mathcal{L}(\widehat{\beta}), \widehat{\Delta}_{(\mathcal{S}_s^* - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s))\cup M^*}\right\rangle + \eta^2\rho^2 C_{\min}^2 \|\widehat{\Delta}_{(\mathcal{S}_s^* - (\widehat{\mathcal{S}}_b \cup \widehat{\mathcal{S}}_s))\cup M^*}\|_F^2 \\
&\qquad\qquad + \eta\left\langle \nabla\mathcal{L}(\widehat{\beta}), \widehat{\Delta}_{\underline{\mathcal{S}}_b^* - (\widehat{\underline{\mathcal{S}}}_b \cup \widehat{\underline{\mathcal{S}}}_s)}\right\rangle + \eta^2\rho^2 C_{\min}^2 \|\widehat{\Delta}_{\underline{\mathcal{S}}_b^* - (\widehat{\underline{\mathcal{S}}}_b \cup \widehat{\underline{\mathcal{S}}}_s)}\|_F^2 \\
&\qquad \leq \eta\Big(\mathcal{L}\left(\beta^*\right) - \mathcal{L}\left(\widehat{\beta}\right)\Big) + \eta^2\rho^2 C_{\min}^2 \|\widehat{\Delta}\|_F^2.
\end{aligned} \tag{6.8}$$

Here, we use the fact that $\nabla\mathcal{L}(\widehat{\beta})$ is zero on the support of $\widehat{\beta}$. Optimizing the RHS over η, we obtain

$$-|(\mathcal{S}_s^* \cup \mathcal{S}_b^*) - (\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_s)|\epsilon_s < -\frac{\left(\mathcal{L}(\beta^*) - \mathcal{L}\left(\widehat{\beta}\right)\right)^2}{4\,\rho^2 C_{\min}^2\,\|\widehat{\Delta}\|_F^2},$$

whence the lemma follows.

□

Lemma 48 (Stopping Error Bound)**.** *When the algorithm stops with parameter $\widehat{\beta}$, we have*

$$\|\widehat{\beta} - \beta^*\|_F \leq \frac{1}{C_{\min}}\left(\frac{\lambda_n}{C_{\min}}\sqrt{|\mathcal{S}_s^* \cup \mathcal{S}_b^*\ \cup\ \widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_b|} + 2\rho\sqrt{|(\mathcal{S}_s^* \cup \mathcal{S}_b^*) - (\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_s)|\epsilon_s}\right). \tag{6.9}$$

Proof. For $\Delta \in \mathbb{R}$, let

$$G(\Delta) = \mathcal{L}\left(\beta^* + \Delta\right) - \mathcal{L}\left(\beta^*\right) - 2\rho C_{\min}\sqrt{|(\mathcal{S}_s^* \cup \mathcal{S}_b^*) - (\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_s)|\epsilon_s}\|\Delta\|_F.$$

It can be seen that $G(0) = 0$, and from the previous lemma, $G(\widehat{\Delta}) \leq 0$. Further, $G(\Delta)$ is sub-homogeneous (over a limited range): $G(t\Delta) \leq tG(\Delta)$ for $t \in [0, 1]$. Thus, for a carefully chosen $r > 0$, if we show that $G(\Delta) > 0$ for all $\Delta \in \{\Delta : \|\Delta\|_2 = r,\ \|\Delta^{(j)}\|_0 \leq s\}$, where, s is the maximum column support size of matrices supported on $\mathcal{S}_s^* \cup \mathcal{S}_b^* \cup \widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_b$, then, it follows that $\|\widehat{\Delta}\|_2 \leq r$. If not, then there would exist some $t \in [0, 1)$ such that $\|t\widehat{\Delta}\| = r$, whence we would arrive at the contradiction

$$0 < G(t\widehat{\Delta}) \leq tG(\widehat{\Delta}) \leq 0. \tag{6.10}$$

Thus, it remains to show that $G(\Delta) > 0$ for all $\Delta \in \{\Delta : \|\Delta\|_2 = r,\ \|\Delta^{(j)}\|_0 \leq s\}$. By restricted strong convexity property of $\mathcal{L}$, we have

$$\begin{aligned}\mathcal{L}(\beta^* + \Delta) - \mathcal{L}(\beta^*) &= \sum_{j=1}^{r}\left(\mathcal{L}(\beta^* + \Delta^{(j)}e_j^T) - \mathcal{L}(\beta^*)\right) \\ &\geq \langle \nabla\mathcal{L}(\beta^*), \Delta\rangle + C_{\min}^2\left\|\Delta\right\|_F^2.\end{aligned}$$

We can establish

$$\begin{aligned}\langle \nabla\mathcal{L}(\beta^*), \Delta\rangle &\geq -\left|\langle \nabla\mathcal{L}(\beta^*), \Delta\rangle\right| \\ &\geq -\left\|\nabla\mathcal{L}(\beta^*)\right\|_\infty\left\|\Delta\right\|_1 = -\lambda_n\left\|\Delta\right\|_1,\end{aligned} \tag{6.11}$$

and hence,

$$\begin{aligned}G(\Delta) &\geq -\lambda_n\|\Delta\|_1 + C_{\min}^2\|\Delta\|_F^2 - 2\rho C_{\min}\sqrt{|(\mathcal{S}_s^* \cup \mathcal{S}_b^*) - (\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_s)|\epsilon_s}\|\Delta\|_F \\ &> \left(-\lambda_n\sqrt{\left|\mathcal{S}_s^* \cup \mathcal{S}_b^* \cup \widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_b\right|} + C_{\min}^2\|\Delta\|_F - 2\rho C_{\min}\sqrt{|(\mathcal{S}_s^* \cup \mathcal{S}_b^*) - (\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_s)|\epsilon_s}\right)\|\Delta\|_F \\ &> 0,\end{aligned} \tag{6.12}$$

if $\|\Delta\|_2 = r$ for

$$r = \frac{1}{C_{\min}}\left(\frac{\lambda_n}{C_{\min}}\sqrt{\left|\mathcal{S}_s^* \cup \mathcal{S}_b^* \cup \widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_b\right|} + 2\rho\sqrt{|(\mathcal{S}_s^* \cup \mathcal{S}_b^*) - (\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_s)|\epsilon_s}\right).$$

This concludes the proof of the lemma.

□

Next, we note that when the algorithm terminates, the backward step with the current parameters has failed to go through. This entails that

$$\begin{aligned} \inf_{i \in \widehat{\underline{\mathcal{S}}}_b} \mathcal{L}(\widehat{\beta} - e_i \widehat{\beta}_i) - \mathcal{L}(\widehat{\beta}) \quad &> \nu \epsilon_b \\ \inf_{(i,j) \in \widehat{\mathcal{S}}_s} \mathcal{L}(\widehat{\beta} - \widehat{\beta}_i^{(j)} e_i e_j^T) - \mathcal{L}(\widehat{\beta}) &> \nu \epsilon_s. \end{aligned} \tag{6.13}$$

The next lemma shows the consequence of this bound.

Lemma 49 (Stopping Backward Step)**.** *When the algorithm stops with parameter $\widehat{\beta}$, we have*

$$\left\| \widehat{\beta}_{(\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_b) - (\mathcal{S}_s^* \cup \mathcal{S}_b^*)} \right\|_F^2 \geq \frac{\nu \epsilon_s}{\rho^2 C_{\min}^2} |(\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_b) - (\mathcal{S}_s^* \cup \mathcal{S}_b^*)|.$$

Proof. We have

$$\begin{aligned} &|(\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_b) - (\mathcal{S}_s^* \cup \mathcal{S}_b^*)| \nu \epsilon_s \\ &\quad \leq |\widehat{\mathcal{S}}_s - (\mathcal{S}_s^* \cup \mathcal{S}_b^*)| \nu \epsilon_s + |\widehat{\mathcal{S}}_b - (\mathcal{S}_s^* \cup \mathcal{S}_b^*)| \nu \epsilon_s \\ &\quad \leq |\widehat{\mathcal{S}}_s - (\mathcal{S}_s^* \cup \mathcal{S}_b^*)| \nu \epsilon_s + |\widehat{M}| \nu \epsilon_s + |\widehat{\underline{\mathcal{S}}}_b - (\underline{\mathcal{S}}_s^* \cup \underline{\mathcal{S}}_b^*)| \nu \epsilon_b \\ &\quad \leq \sum_{(i,j) \in (\widehat{\mathcal{S}}_s - (\mathcal{S}_s^* \cup \mathcal{S}_b^*)) \cup \widehat{M}} \left(\mathcal{L}(\widehat{\beta} - \widehat{\beta}_i^{(j)} e_i e_j^T) - \mathcal{L}(\widehat{\beta}) \right) + \sum_{i \in \widehat{\underline{\mathcal{S}}}_b - (\underline{\mathcal{S}}_s^* \cup \underline{\mathcal{S}}_b^*)} \left(\mathcal{L}(\widehat{\beta} - e_i \widehat{\beta}_i) - \mathcal{L}(\widehat{\beta}) \right) \\ &\quad \leq \underbrace{\left\langle \nabla \mathcal{L}(\widehat{\beta}), \widehat{\beta}_{(\widehat{\mathcal{S}}_s - (\mathcal{S}_s^* \cup \mathcal{S}_b^*)) \cup \widehat{M}} \right\rangle}_{0} + \rho^2 C_{\min}^2 \|\widehat{\beta}_{(\widehat{\mathcal{S}}_s - (\mathcal{S}_s^* \cup \mathcal{S}_b^*)) \cup \widehat{M}}\|_F^2 \\ &\qquad\qquad + \underbrace{\left\langle \nabla \mathcal{L}(\widehat{\beta}), \widehat{\beta}_{\widehat{\underline{\mathcal{S}}}_b - (\underline{\mathcal{S}}_s^* \cup \underline{\mathcal{S}}_b^*)} \right\rangle}_{0} + \rho^2 C_{\min}^2 \|\widehat{\beta}_{\widehat{\underline{\mathcal{S}}}_b - (\underline{\mathcal{S}}_s^* \cup \underline{\mathcal{S}}_b^*)}\|_F^2 \\ &\quad = \rho^2 C_{\min}^2 \left\| \widehat{\beta}_{(\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_b) - (\mathcal{S}_s^* \cup \mathcal{S}_b^*)} \right\|_F^2, \end{aligned} \tag{6.14}$$

where, the second inequality uses the fact that $[\nabla \mathcal{L}(\widehat{\beta})]_{\widehat{\mathcal{S}}_s \cup \widehat{\mathcal{S}}_b} = 0$. Substituting (6.13) above, the lemma follows. □

6.6 Lemmas on the Stopping Size

Let $\mathcal{S}^*(j) = \{i : (i,j) \in \mathcal{S}_s^* \cup \mathcal{S}_b^*\}$ and notice that this is larger than the support of the j^{th} column of β^* (because there might be a feature shared across many, but not all, tasks and that feature is on the support set of the block-sparse matrix). Also, for $k = k_s + k_b - 1$, let $\widehat{\mathcal{S}}^{(k_s,k_b)}(j) = \{i : (i,j) \in \widehat{\mathcal{S}}_s^{(k_s)} \cup \widehat{\mathcal{S}}_b^{(k_b)}\}$ be the support of the j^{th} column of our current estimation and let $s_j = |\widehat{\mathcal{S}}^{(k_s,k_b)}(j)|$.

Lemma 50. *If $\epsilon_s > \frac{\lambda_n^2 \rho^2}{C_{\min}^2 f(\eta)}$ for some $\eta \geq 2 + \frac{4r\epsilon_s\rho^4(\rho^4-\rho^2+2)}{\epsilon_b\nu}$ and REP $\left(\eta s_j^*\right)$ holds, then the algorithm stops with $s_j \leq (\eta-1)s_j^*$ for all $j \in \{1,2,\ldots,r\}$.*

Proof. Consider the first time the algorithm reaches $s_j = (\eta-1)s_j^* + 1$. By Lemmas 53 and 52, we have

$$\begin{aligned}
\sqrt{\frac{\epsilon_b\nu}{r\epsilon_s}}\sqrt{\frac{s_j - 1 - s_j^*}{s_j - 1}} &\leq \sqrt{\frac{\nu\epsilon_b}{r\epsilon_s}}\sqrt{\frac{|\widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j) - \mathcal{S}^*(j)|}{|\widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j) \cup \mathcal{S}^*(j)|}} \\
&\leq \frac{\lambda_n\rho}{C_{\min}\sqrt{\epsilon_s}} + \frac{\rho^3\sqrt{2(\rho^2-1)}}{\sqrt{|\widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j) \cup \mathcal{S}^*(j)|}} \\
&\qquad + 2\rho^2\sqrt{\frac{|\mathcal{S}^*(j) - \widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j)|}{|\widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j) \cup \mathcal{S}^*(j)|}} \\
&\leq \frac{\lambda_n\rho}{C_{\min}\sqrt{\epsilon_s}} + \frac{\rho^3\sqrt{2(\rho^2-1)}}{\sqrt{s_j - 1}} + 2\rho^2\sqrt{\frac{s_j^*}{s_j + s_j^* - 1}}.
\end{aligned}$$

Hence, we get

$$f(\eta) := \frac{\sqrt{\frac{\epsilon_b\nu}{r\epsilon_s}(\eta-2)} - \sqrt{\frac{2\rho^6(\rho^2-1)}{s_j^*}}}{\sqrt{\eta-1}} - \frac{2\rho^2}{\sqrt{\eta}} \leq \frac{\lambda_n\rho}{C_{\min}\sqrt{\epsilon_s}}.$$

For $\eta \geq 2 + \frac{4r\epsilon_s\rho^4(\rho^4-\rho^2+2)}{\epsilon_b\nu}$, the LHS is positive and we arrive to a contradiction with the assumption on ϵ_s.

□

Lemma 51 (General Forward Step)**.** *For any $j \in \{1, 2, \ldots, r\}$, the first time the algorithm reaches a (column) support size of s_j at the beginning of the forward step, we have*

$$\left|\mathcal{L}\left(\beta^{*(j)}e_j^T\right) - \mathcal{L}\left(\widehat{\beta}^{(j)}(k-1)e_j^T\right)\right| \leq 2\rho C_{\min}\sqrt{\left|\mathcal{S}^*(j) - \widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j)\right| \mu_s^{(k)}\epsilon_s}\left\|\beta^{*(j)} - \widehat{\beta}^{(j)}(k-1)\right\|_2.$$

Proof. According to the forward step, we have

$$\mathcal{L}\left(\widehat{\beta}(k-1)\right) - \inf_{(i,j)\notin\widehat{\mathcal{S}}_s^{(k_s-1)};\,\gamma\in\mathbb{R}} \mathcal{L}\left(\widehat{\beta}(k-1) + \gamma e_i e_j^T\right) = \mu_s^{(k)}\epsilon_s.$$

Since the loss function is separable with respect to the columns of β, for any fixed $j \in \{1, \ldots, r\}$ we have

$$\mathcal{L}\left(\widehat{\beta}^{(j)}(k-1)e_j^T\right) - \inf_{i:(i,j)\notin\widehat{\mathcal{S}}_s^{(k_s-1)};\,\gamma\in\mathbb{R}} \mathcal{L}\left(\widehat{\beta}^{(j)}(k-1)e_j^T + \gamma e_i e_j^T\right) \leq \mu_s^{(k)}\epsilon_s.$$

Similar to (6.8), for any $\eta \in \mathbb{R}$, we have

$$-\left|\mathcal{S}^*(j) - \widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j)\right| \mu_s^{(k)}\epsilon_s \leq \eta\left(\mathcal{L}\left(\beta^{*(j)}e_j^T\right) - \mathcal{L}\left(\widehat{\beta}^{(j)}(k-1)e_j^T\right)\right) + \eta^2\rho^2 C_{\min}^2\left\|\beta^{*(j)} - \widehat{\beta}^{(j)}(k-1)\right\|_2^2.$$

Optimizing the RHS over η, we obtain

$$\left|\mathcal{S}^*(j) - \widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j)\right| \mu_s^{(k)}\epsilon_s \geq \frac{\left(\mathcal{L}\left(\beta^{*(j)}e_j^T\right) - \mathcal{L}\left(\widehat{\beta}^{(j)}(k-1)e_j^T\right)\right)^2}{4\rho^2 C_{\min}^2\left\|\beta^{*(j)} - \widehat{\beta}^{(j)}(k-1)\right\|_2^2}.$$

This concludes the proof of the lemma.

□

Lemma 52 (General Error Bound). *For any $j \in \{1, 2, \ldots, r\}$, the first time the algorithm reaches a (column) support size of s_j at the beginning of the forward step, we have*

$$\left\|\beta^{*(j)} - \widehat{\beta}^{(j)}(k-1)\right\|_2 \leq \frac{\lambda_n}{C_{\min}^2}\sqrt{\left|\mathcal{S}^*(j) \cup \widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j)\right|} + \frac{2\rho}{C_{\min}}\sqrt{\left|\mathcal{S}_s^*(j) - \widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j)\right| \mu_s^{(k)}\epsilon_s}.$$

Proof. The proof is identical to the proof of lemma 48 and is omitted. □

Lemma 53 (General Backward Step). *For any $j \in \{1, 2, \ldots, r\}$, the first time the algorithm reaches a (column) support size of s_j at the beginning of the forward step, if $s_j > s_j^* + \frac{2r\rho^6(\rho^2-1)}{\nu}$, then*

$$\left\|\widehat{\beta}^{(j)}_{\widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j)-\mathcal{S}^*(j)}(k-1)\right\|_2^2 \geq \left(\frac{\sqrt{\left|\widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j) - \mathcal{S}^*(j)\right|\nu}}{\rho C_{\min}}\sqrt{\frac{\epsilon_b}{r\epsilon_s}} - \frac{\rho^2\sqrt{2(\rho^2-1)}}{C_{\min}}\right)^2 \mu^{(k)}\epsilon_s.$$

Proof. Under the assumption of the lemma, the immediate previous backward step has not gone through and hence,

$$\inf_{(i,j)\in\widehat{\mathcal{S}}_s^{(k_s-1)}} \mathcal{L}\left(\widehat{\beta}(k) - \widehat{\beta}_i^{(j)}(k)e_ie_j^T\right) - \mathcal{L}\left(\widehat{\beta}(k)\right) \geq \nu\mu^{(k)}\epsilon_s$$

$$\inf_{i\in\widehat{\mathcal{S}}_b^{(k_b-1)}} \mathcal{L}\left(\widehat{\beta}(k) - e_i\widehat{\beta}_i(k)\right) - \mathcal{L}\left(\widehat{\beta}(k)\right) \geq \nu\mu^{(k)}\epsilon_b.$$

Since the loss function is separable with respect to the columns of β, for a fixed $j \in \{1, 2, \ldots, r\}$, we have

$$\inf_{i:(i,j)\in\widehat{\mathcal{S}}_s^{(k_s-1)}} \mathcal{L}\left(\widehat{\beta}^{(j)}(k)e_j^T - \widehat{\beta}_i^{(j)}(k)e_ie_j^T\right) - \mathcal{L}\left(\widehat{\beta}^{(j)}(k)e_j^T\right) \geq \nu\mu^{(k)}\epsilon_s.$$

Consequently, similar to (6.14), we can show that

$$\begin{aligned}
&\left|\widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j) - \mathcal{S}^*(j)\right| \nu\mu^{(k)}\frac{\epsilon_b}{r} \\
&\quad\leq \left|\widehat{\mathcal{S}}_s^{(k_s-1)}(j) - \mathcal{S}^*(j)\right| \nu\mu^{(k)}\epsilon_s + \left|\widehat{\mathcal{S}}_b^{(k_b-1)}(j) - (\mathcal{S}^*(j)\cup\widehat{\mathcal{S}}_s^{(k_s-1)}(j))\right| \nu\mu^{(k)}\frac{\epsilon_b}{r} \\
&\quad\leq \rho^2 C_{\min}^2 \left\|\widehat{\beta}^{(j)}_{\widehat{\mathcal{S}}_s^{(k_s-1)}(j)-\mathcal{S}^*(j)}(k)\right\|_2^2 + \rho^2 C_{\min}^2 \left\|\widehat{\beta}_{\widehat{\mathcal{S}}_b^{(k_b-1)}(j)-(\mathcal{S}^*(j)\cup\widehat{\mathcal{S}}_s^{(k_s-1)}(j))}(k)\right\|_{2,\infty}^2 \\
&\quad\leq \rho^2 C_{\min}^2 \left(\left\|\widehat{\beta}^{(j)}_{\widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j)-\mathcal{S}^*(j)}(k-1)\right\|_2 + \left\|\Delta^{(k)}\right\|_2\right)^2,
\end{aligned}$$

where, $\Delta^{(k)} = \widehat{\beta}^{(j)}_{\widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j)}(k) - \widehat{\beta}^{(j)}(k-1)$. This entails that

$$\left(\frac{\sqrt{\left|\widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j) - \mathcal{S}^*(j)\right| \nu\mu_s^{(k)}\frac{\epsilon_b}{r}}}{\rho C_{\min}} - \left\|\Delta^{(k)}\right\|_F\right)^2 \leq \left\|\widehat{\beta}_{\widehat{\mathcal{S}}^{(k_s-1,k_b-1)}(j)-\mathcal{S}^*(j)}(k-1)\right\|_F^2.$$

Thus, it suffices to show that $\left\|\Delta^{(k)}\right\|_F \leq \frac{\rho^2}{C_{\min}}\sqrt{2(\rho^2-1)\mu_s^{(k)}\epsilon_s}$ since $\mu_s^{(k)} \leq \mu^{(k)}$. Notice that by our assumption on the size of the support, the first term is always larger than the second provided we can show this inequality. There are two cases: (a) if we added a single element in the previous step for which we show the above inequality, and (b) if we added a row in the previous step for which we show $\left\|\Delta^{(k)}\right\|_F \leq \frac{\rho^2}{C_{\min}}\sqrt{2(\rho^2-1)\mu_b^{(k)}\frac{\epsilon_b}{r}}$. Since $\frac{\epsilon_b}{r} \leq \epsilon_s$ and $\mu_b^{(k)} \leq \mu^{(k)}$, the result follows. We prove (a) and omit the proof of (b) since it is identical.

We drop the super- and sub-script j for the ease of the notation in the rest of the proof. From the forward step, we have

$$\mathcal{L}\left(\widehat{\beta}(k-1)\right) - \inf_{(i,j)\notin\widehat{\mathcal{S}}_s^{(k-1)},\gamma\in\mathbb{R}} \mathcal{L}\left(\widehat{\beta}(k-1) + \gamma e_i e_j^T\right) = \mu_s^{(k)}\epsilon_s.$$

Let $(i_*, j_*, \gamma_* \neq 0)$ be the optimizer of the equation above. Now, we have

$$\begin{aligned}
C_{\min}^2\left\|\Delta^{(k)}\right\|_2^2 &\leq \mathcal{L}\left(\widehat{\beta}(k)_{\widehat{\mathcal{S}}_s^{(k-1)}\cup\widehat{\mathcal{S}}_b^{(k-1)}}\right) - \mathcal{L}\left(\widehat{\beta}(k-1)\right) \\
&\leq \mathcal{L}\left(\widehat{\beta}(k)_{\widehat{\mathcal{S}}_s^{(k-1)}\cup\widehat{\mathcal{S}}_b^{(k-1)}}\right) - \mathcal{L}\left(\widehat{\beta}(k)\right) + \mathcal{L}\left(\widehat{\beta}(k)\right) - \mathcal{L}\left(\widehat{\beta}(k-1)\right) \\
&\leq \rho^2 C_{\min}^2\left|\widehat{\beta}_{i_*}^{(j_*)}(k)\right|^2 - C_{\min}^2\left\|\Delta^{(k)}\right\|_2^2 - C_{\min}^2\left|\widehat{\beta}_{i_*}^{(j_*)}(k)\right|^2.
\end{aligned}$$

Hence, $\left\|\Delta^{(k)}\right\|_2^2 \leq \frac{\rho^2-1}{2}\left|\widehat{\beta}_{i_*}^{(j_*)}(k)\right|^2$ and we only need to show that $\left|\widehat{\beta}_{i_*}^{(j_*)}(k)\right| \leq \frac{2\rho^2\sqrt{\mu_s^{(k)}\epsilon_s}}{C_{\min}}$. Since $\left|\widehat{\beta}_{i_*}^{(j_*)}(k)\right| \leq \left|\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right| + |\gamma_*|$, we can equivalently control the latter two terms. First, by forward step construction, $C_{\min}^2|\gamma^*|^2 \leq \mathcal{L}\left(\widehat{\beta}(k-1)\right) - \mathcal{L}\left(\widehat{\beta}(k-1)+\gamma_* e_{i_*}e_{j_*}^T\right) = \mu_s^{(k)}\epsilon_s$ and hence $|\gamma^*| \leq \frac{\sqrt{\mu_s^{(k)}\epsilon_s}}{C_{\min}}$. Second, we claim that $\left|\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right| \leq (2\rho^2-1)|\gamma_*|$ and we are done.

In contrary, suppose $\left|\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right|^2 > (2\rho^2-1)^2|\gamma_*|^2 \geq \rho^2|\gamma_*|^2$. We have

$$\begin{aligned}
C_{\min}^2\left|\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right|^2 &> \rho^2 C_{\min}^2|\gamma_*|^2 \\
&\geq \mathcal{L}\left(\widehat{\beta}(k)-\gamma_* e_{i_*}e_{j_*}^T\right) - \mathcal{L}\left(\widehat{\beta}(k)\right) \\
&\geq \mathcal{L}\left(\widehat{\beta}(k)-\gamma_* e_{i_*}e_{j_*}^T\right) - \mathcal{L}\left(\widehat{\beta}(k-1)\right) + \mathcal{L}\left(\widehat{\beta}(k-1)\right) - \mathcal{L}\left(\widehat{\beta}(k)\right) \\
&\geq C_{\min}^2\left\|\Delta^{(k)}\right\|_2^2 + C_{\min}^2\left|\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right|^2 + \nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\left(\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right) \\
&\quad + C_{\min}^2\left\|\Delta^{(k)}\right\|_2^2 + C_{\min}^2\left|\widehat{\beta}_{i_*}^{(j_*)}(k)\right|^2.
\end{aligned}$$

This is a contradiction provided $C_{\min}^2\left|\widehat{\beta}_{i_*}^{(j_*)}(k)\right|^2 + \nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\left(\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right) \geq 0$. Later in the proof, we will show that $\text{Sign}\left(\nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\right) = -\text{Sign}(\gamma_*)$ and that $2C_{\min}^2|\gamma_*| \leq \left|\nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\right| \leq 2\rho^2 C_{\min}^2|\gamma_*|$. With these, if $\frac{\widehat{\beta}_{i_*}^{(j_*)}(k)}{\gamma_*} \leq 1$, we have $\nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\left(\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right) \geq 0$ and the claim follows. Otherwise, we have $\left|\widehat{\beta}_{i_*}^{(j_*)}(k)\right| \geq \left|\widehat{\beta}_{i_*}^{(j_*)}(k)\right| - |\gamma_*| = \left|\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right|$ so that $\left|\widehat{\beta}_{i_*}^{(j_*)}(k)\right| \geq 2\rho^2|\gamma_*|$ and hence,

$$\begin{aligned}
&C_{\min}^2\left|\widehat{\beta}_{i_*}^{(j_*)}(k)\right|^2 + \nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\left(\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right) \\
&\qquad \geq 2\rho^2 C_{\min}^2|\gamma_*|\left|\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right| - 2\rho^2 C_{\min}^2|\gamma_*|\left|\widehat{\beta}_{i_*}^{(j_*)}(k)-\gamma_*\right| \\
&\qquad = 0.
\end{aligned}$$

To get the claimed properties of $\nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)$, note that

$$\begin{aligned} C_{\min}^2 \left|\gamma_*\right|^2 &\leq \mathcal{L}\left(\widehat{\beta}(k-1)\right) - \mathcal{L}\left(\widehat{\beta}(k-1) + \gamma_* e_{i_*} e_{j_*}^T\right) \\ &\leq -C_{\min}^2 \left|\gamma_*\right|^2 - \nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\gamma_* \,, \end{aligned}$$

and hence $\text{Sign}\left(\nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\right) = -\text{Sign}\left(\gamma_*\right)$ and $2C_{\min}^2|\gamma_*| \leq \left|\nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\right|$. Also, we can establish

$$\begin{aligned} \rho^2 C_{\min}^2 \left|\gamma_*\right|^2 &\geq \mathcal{L}\left(\widehat{\beta}(k-1)\right) - \mathcal{L}\left(\widehat{\beta}(k-1) + \gamma_* e_{i_*} e_{j_*}^T\right) \\ &\geq -\rho^2 C_{\min}^2 \left|\gamma_*\right|^2 - \nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\gamma_* \,. \end{aligned}$$

Since $-\nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\gamma_* \geq 0$, we can conclude that $\left|\nabla_{(i_*,j_*)}\mathcal{L}\left(\widehat{\beta}(k-1)\right)\right| \leq 2\rho^2 C_{\min}^2|\gamma_*|$. This concludes the proof of the lemma.

□

Chapter 7

Conclusion

The work thus far was dedicated to study of the behavior of the dirty models in a high-dimensional regime. We studied two major classes of dirty models:

- **Sparse plus Block Sparse**: We showed that in multiple linear regression problem, dirty model outperforms clean models in the sense that with dirty models, we require less number of observations for structure recovery. Simulations support the theoretical results. We observed results for algorithms based on both convex optimization and greedy approach.
- **Sparse plus Low-Rank**: We formulated the graph clustering problem as a sparse plus low-rank dirty model of the adjacency matrix. Then, we derived sufficient conditions under which a convex optimization process can recover the clusters exactly using two different convex algorithms of nuclear norm and max-norm minimizations. We verified the theoretical result by simulation.

  We also considered the problem of learning dependency graph of a partially observed time-series. We showed that this problem can also be formulated as an sparse plus low-rank dirty model, where sparse matrix captures the effect of observed variables and low-rank part captures the effect of latent variables. We provided conditions under which this algorithm succeeds and verified our results on stock market data.

I would like to continue my research on different paths as listed below.

7.1 Flexible and Robust High-dimensional Statistics

Following up on the discussion of robustness and flexibility in the introduction, I would like to explore three different paths under this topic:

(1) I want to study more sophisticated, and perhaps more complicated, high-dimensional structures, like non-linear mixture of simple structures, that fit the real applications better. Some of my previous works along these lines are (a) modeling the multi-task problem as a sparse plus block-sparse structure recovery, (b) recovery of a low-rank matrix when some of the entries are corrupted and some of the entries are missing, (c) modeling the graph clustering problem as a sparse plus low-rank structure recovery. In all these cases, I provide a recovery algorithm plus theoretical guarantee.

(2) I want to investigate the effect of time-dependency on the high-dimensional learning algorithms. In most research papers, there is an independence assumption of the observations. However, in reality, often times, the observations are correlated over the time. Since independent observations carry more information than correlated observations, it is not clear if the existing algorithms can perform well under correlation assumption. Thus far, I have considered the problem of learning a network of stochastic differential equations when some of the functions are not observed. I have shown that this problem can be modeled by a flexible super-position of a sparse and a low-rank model.

(3) I want to develop a general framework for flexible high-dimensional statistical problems. Most of the work in this area seems to follow same principles in general, however, the different (sufficient) assumptions of different works and different techniques make the comparison of algorithms hard. I believe a unified framework and finding the minimum requirement for these algorithms can significantly improve our understanding of high-dimensional statistics.

7.2 Graphical Data Modeling in High-dimensions

Dealing with data, leveraging the concept of graph as a platform for dependency illustration, visualization, representation, management, database design, etc, becomes more and more popular. Finding the right graph structure for a certain application given (potentially noisy) data is a challenge. For high-dimensional data, the problems become even harder because of potential inconsistencies. One of the popular use of graphs in data analysis is using the graph to show the conditional dependencies of different variables/features, often referred to as graphical models. Given some realization of a number of random variables, we are interested in learning how these variables are correlated in high-dimensional setting (i.e., when we have too many random variables, but very few realizations). This correlation can be restricted to pairwise correlation of each pair of random variables or can be extended to higher-order correlations between any sub-set of the random variables. So far, I have investigated the problem of learning pairwise and higher-order graphical models for general discrete random variables.

I am interested in conducting research in two different areas under this topic:

(1) I want to study the problem of learning higher-order dependency graphical models with low-complexity algorithms. Since there are exponentially many sub-sets of a set of random variables, all existing algorithms for higher-order graphical models are exponential in the number of variables. However, in a high-dimensional setting, this is not acceptable.

(2) I want to develop efficient algorithms for learning directed (causal) graphical models in high-dimensional setting. Directed graphical models assume that one variable *causes* another variable and hence, they model the dependencies as a directed graph. A popular example of such graphs in machine learning is decision trees.

7.3 Information Theoretic Limits of High-dimensional Statistics

Often times the estimation/prediction problems in high-dimensional statistics are NP-hard and researchers consider a convex relaxation of the original problem. While there exist many algorithms with guarantees under certain set of assumptions, it has not been well studied that what the necessary conditions for the estimation/prediction are. Finding these necessary bounds helps us to have a better understanding of the performance of different methods.

I am interested in following two different paths under this topic:

(1) I like to develop low-complexity learning algorithms that can be applied to high-dimensional datasets without convexification. While most of the existing algorithms for solving convex optimization programs are polynomial-time algorithms, they are not scalable to high-dimensional datasets. Moreover, most of them are solving a convexified version of the original problem and hence, limited in the nature. As a primary work, I have proposed a greedy algorithm for learning sparse structures without convexification. I have shown that not only the algorithm performs better than those who solve convex programs, but also, it can be applied to broader instances of the problem.

(2) I want to work on information theoretic bounds for estimation/prediction in high-dimensional setting when the data has some structure. Such bounds show that how close our algorithms are to the optimal and more importantly, how much we lose by convexification.

Bibliography

[1] A. Agarwal, S. Negahban, and M. Wainwright. Convergence rates of gradient methods for high-dimensional statistical recovery. In *NIPS*, 2010.

[2] N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. In *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 457–466, 1998.

[3] N. Alon and A. Naor. Approximating the cut-norm via grothendieck's inequality. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 72–80. ACM, 2004.

[4] N. Alon and A. Noar. Approximating the cut-norm via grothendieck's inequality. *SIAM Journal on Computing*, 35, 2006.

[5] B. Ames and S. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Mathematical Programming*, 129(1):69–89, 2011.

[6] E.M. Azoff. *Neural Network Time Series Forecasting of Financial Markets*. John Wiley & Sons, Inc., 1994.

[7] F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[8] F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Neural Information Processing Systems (NIPS)*, 2008.

[9] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[10] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. In *Proceedings of the 43rd Symposium on Foundations of Computer Science*, 2002.

[11] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. In *Proceedings of the 43rd Symposium on Foundations of Computer Science*, 2002.

[12] Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics, Oxford University Press*, 20:2493–2503, 2004.

[13] R. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

[14] H. Becker. A survey of correlation clustering. Available online at http://www1.cs.columbia.edu/ hila/clustering.pdf, 2005.

[15] J. Bento, M. Ibrahimi, and A. Montanari. Learning networks of stochastic equations. In *NIPS*, 2010.

[16] A. Berman and N. Shaked-Monderer. *Completely Positive Matrices*. World Scientific Publication, 2003.

[17] M. Bern and D. Eppstein. *Approximation Algorithms for Geometric Problems*. In Approximation Algorithms for NP-Hard Problems, edited by D. S. Hochbaum, Boston: PWS Publishing Company, 1996.

[18] B.L. Bowerman and R.T. O'Connell. *Forecasting and time series: An applied approach*. Duxbury Press, 1993.

[19] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time-series Analysis: Forecasting and Control*. John Wiley & Sons, Inc., 1990.

[20] S. Burer and C. Choi. Computational enhancements in low-rank semidefinite programming. *Optimization Methods and Software*, 21:493–512, 2006.

[21] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? Technical report, Stanford University, CA, 2009.

[22] E. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 2009.

[23] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? In *Available at arXiv:0912.3599*, 2009.

[24] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58:1–37, 2011.

[25] E. J. Candes and Y. Plan. Matrix completion with noise. In *IEEE Proceedings*, volume 98, pages 925 – 936, 2010.

[26] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[27] P. Chan, M. Schlag, and J. Zien. Spectral k-way ratio cut partitioning. *IEEE Trans. CAD-Integrated Circuits and Systems*, 13:1088–1096, 1994.

[28] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. arXiv:1008.1290, August 2010.

[29] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. In *Available at arXiv:1008.1290*, 2010.

[30] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 2011.

[31] V. Chandrasekaran, S. Sanghavi, S. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[32] M. Charikar, V. Guruswami, and A. Wirth. Culstering with qualitative information. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, 2003.

[33] C. Chatfield. *Time-series Forecasting*. Chapman & Hall, 2000.

[34] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. In *International Symposium on Information Theory*, pages 2313 – 2317, 2011.

[35] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. In *ISIT*, 2011.

[36] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. In *ISIT*, 2011.

[37] T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris. A robust and scalable clustering algorithm for mixed type attributes in large database environment. *Proceedings of the international conference on Knowledge discovery and data mining*, 2001.

[38] J. H. Cochrane. *Time Series for Macroeconomics and Finance*. University of Chicago, 2005.

[39] A. Condon and R.M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures & Algorithms*, 2001.

[40] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and banach spaces. *Handbook of Banach Spaces*, 1:317–336, 2001.

[41] D. Defays. An efficient algorithm for a complete link method. *The Computer Journal (British Computer Society)*, 20:364–366, 1977.

[42] E. D. Demaine, N. Immorlica, D. Emmanuel, and A. Fiat. Correlation clustering in general weighted graphs. *SIAM special issue on approximation and online algorithms*, 2005.

[43] I. Dhillon, Y. Guan, and B. Kulis. A unified view of kernel k-means, spectral clustering and graph cuts. *UTCS Technical Report TR-04-25, University of Texas at Austin*, 2005.

[44] R. P.W. Duin. Department of Applied Physics, Delft University of Technology, Delft, The Netherlands, 2002.

[45] D. Emmanuel and A. Fiat. Correlation clustering minimizing disagreements on arbitrary weighted graphs. In *Proceedings of the 11th Annual European Symposium on Algorithms*, 2003.

[46] D. Emmanuel and N. Immorlica. Correlation clustering with partial information. In *Proceedings of the 6th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems*, 2003.

[47] M. Ester, H. Kriegel, and X. Xu. A database interface for clustering in large spatial databases. 1995.

[48] B. Everitt. *Cluster Analysis*. New York: Halsted Press, 1980.

[49] M. Fazel, T. K. Pong, D. Sun, and P. Tseng. Hankel matrix rank minimization with applications in system identification and realization. University Of Washington, Seattle, WA, 2011.

[50] X. Z. Fern and C. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *20th International Conference on Machine learning (ICML)*, 2003.

[51] R. A. Fisher. Theory of statistical estimation. In *Proceedings of Cambridge Philosophy Society*, volume 22, pages 700–725, 1925.

[52] G.W. Flake, R.E. Tarjan, and K. Tsioutsiouliklis. Graph clustering and minimum cut trees. *Internet Mathematics*, 2004.

[53] J. Friedman, T. Hastie, and R. Tibshirani. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2010.

[54] S. Geršgorin. Uber die abgrenzung der eigenwerte einer matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, 7:749–754, 1931.

[55] D.T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007.

[56] M. Goemans and D. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of ACM*, 42, 1995.

[57] L.J. Gray and D.G. Wilson. Nonnegative factorization of positive semidefinite nonnegative matrices. *Linear Algebra and its Applications*, 31:119–127, 1980.

[58] D. Gross. Recovering low-rank matrices from few coefficients in any basis. Available on arXiv:0910.1879v4, 2009.

[59] M. Handcock, A. Raftery, and J. Tantrum. Model-based clustering for social networks. *Statistics in Society*, 2007.

[60] X. He and P. Niyogi. Locality preserving projections. In *NIPS*, 2003.

[61] D. Higham. Modeling and simulating chemical reactions. *SIAM Review*, 50:347–368, 2008.

[62] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[63] Daniel Hsu, Sham Kakade, and Tong Zhang. Robust matrix decomposition with outliers. *Available at arXiv:1011.1518*, 2010.

[64] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *International Conference on Machine Learning (ICML)*, 2009.

[65] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1981.

[66] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[67] A. Jalali, Y. Chen, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. In *ICML*, 2011.

[68] A. Jalali, C. Johnson, and P. Ravikumar. On learning discrete graphical models using greedy methods. In *NIPS*, 2011.

[69] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *NIPS*, 2010.

[70] A. Jalali, P. Ravikumar, V. Vasuki, and S. Sanghavi. On learning discrete graphical models using group-sparse regularization. In *Inter. Conf. on AI and Statistics (AISTATS) 14*, 2011.

[71] Hiriart-Urruty J.B. and Lembarechal C. *Convex Analysis and Minimization Algorithms I.* Springer-Verlag, Netherland, 1991.

[72] M. I. Jordan. *Learning in Graphical Models.* Kluwer Academic Publishers, Netherland, 1998.

[73] R. Kannan, S. Vempala, and A. Vetta. On clusterings - good, bad and spectral. In *IEEE Symposium on Foundations of Computer Science*, 2000.

[74] B. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, 49(2):291–307, 1970.

[75] B.W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307, 1970.

[76] K. Kim. Financial time series forecasting using support vector machines. *Elsevier Neurocomputing*, 55:307–319, 2003.

[77] B. Krishnamurthy. An improved min-cut algorithm for partitioning vlsi networks. *IEEE Transactions on Computers*, 1984.

[78] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):957–968, 2005.

[79] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *The Journal of Machine Learning Research*, 10:777–801, 2009.

[80] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28:1303–1338, 1998.

[81] N. D. Lawrence, M. Girolami, M. Rattray, and G. Sanguinetti. *Learning and Inference in Computational Systems Biology.* MIT Press, 2010.

[82] J. Lee, B. Recht, R. Salakhutdinov, N. Srebro, and J.A. Tropp. Practical large-scale optimization for max-norm regularization. In *NIPS*, 2010.

[83] T. Lee and A. Shraibman. A direct product theorem for discrepancy. In *Proceedings of the IEEE 23rd Annual Conference on Computational Complexity*, 2008.

[84] Z. Lin, M. Chen, L. Wu, and Y. Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *UIUC Technical Report UILU-ENG-09-2215*, 2009.

[85] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. MA. Fast convex optimization algorithm for exact recovery of a corrupted low-rank matrix. *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2009.

[86] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. In *UIUC Technical Report UILU-ENG-09-2214*, 2009.

[87] N. Linial, S. Mendelson, G. Schechtman, and A. Shraibman. Complexity measures of sign matrices. *Combinatorica*, 27(4):439–463, 2007.

[88] H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *26th International Conference on Machine Learning (ICML)*, 2009.

[89] K. Lounici, A. B. Tsybakov, M. Pontil, and S. A. van de Geer. Taking advantage of sparsity in multi-task learning. In *22nd Conference On Learning Theory (COLT)*, 2009.

[90] S. Mancoridis, B. Mitchell, C. Rorres, Y. Chen, and E. Gansner. Using automatic clustering to produce high-level system organizations of source code. In *Proceedings of the 6th International Workshop on Program Comprehension*, 1998.

[91] P. Marchal. Constructing a sequence of random walks strongly converging to brownian motion. In *Discrete Mathematics and Theoretical Computer Science Proceedings*, pages 181–190, 2003.

[92] R. Mathias. Spectral perturbation bounds for positive definite matrices. *SIAM J. on Matrix Analysis and Applications*, 14:959–980, 1997.

[93] C. Mathieu and W. Schudy. Correlation clustering with noisy input. In *Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, 2010.

[94] L. Meier, S. van de Geer, and P. Buhlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society*, 70:53–71, 2008.

[95] M. Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98:873–895, 2007.

[96] M. Meilă and J. Shi. Learning segmentation by random walks. In *NIPS*, 2001.

[97] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[98] N. Mishra, I. Stanton R. Schreiber, and R. E. Tarjan. Clustering social networks. *Algorithms and Models for Web-Graph, Springer*, 2007.

[99] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Neural Information Processing Systems (NIPS) 22*, 2009.

[100] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Arxiv*, 2010.

[101] S. Negahban and M. J. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[102] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. In *ICML*, 2010.

[103] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2002.

[104] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 2010.

[105] S. Oymak and B. Hassibi. Finding dense clusters via low rank + sparse decomposition. Available on arXiv:1104.5186v1, 2011.

[106] J. Peltonen and S. Kaski. Discriminative components of data. *IEEE Transactions on Neural Networks*, 2004.

[107] J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1):186–205, 2007.

[108] S. Perkins and J. Theiler. Online feature selection using grafting. In *ICML*, 2003.

[109] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society, Series B*, 2009.

[110] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society, Series B*, 2009.

[111] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 2009.

[112] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

[113] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Technical Report 767, UC Berkeley, Department of Statistics*, 2008.

[114] B. Recht. A Simpler Approach to Matrix Completion. *Arxiv preprint arXiv:0910.0651*, 2009.

[115] B. Recht, M. Fazel, and P. Parillo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. Available on arXiv:0706.4138v1, 2009.

[116] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. In *Allerton Conference, Allerton House, Illinois*, 2007.

[117] R. T. Rockafellar. *Convex Analysis*. Prenticeton University Press; Princeton, NJ, 1970.

[118] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[119] E. Schaeffer. Graph clustering. *Computer Science Review*, 2007.

[120] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[121] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[122] S. E. Shreve. *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer, 2004.

[123] R. Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal (British Computer Society)*, 16(1):30–34, 1973.

[124] P. M. A. Sneath and R. R. Sokal. *Numerical Taxonomy*. Freeman, San Francisco CA, 1973.

[125] N. Srebro. *Learning with Matrix Factorizations*. PhD thesis, Massachusetts Institute of Technology, Boston, MA, 2004.

[126] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. In *NIPS*, 2005.

[127] H. Steinhaus. Sur la division des corps matériels en parties. *Bulletin de l'Acadbmie Polonaise des Sciences*, 4:801–804, 1957.

[128] G. W. Stewart. On the perturbation of schur complements in positive semidefinite matrices. *Technical Report, University of Maryland, College Park*, 1995.

[129] C. Swamy. Correlation clustering: maximizing agreements via semidefinite programming. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2004.

[130] V. N. Temlyakov. Greedy approximation. *Acta Numerica*, 17:235–409, 2008.

[131] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[132] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, Series B*, 58:267–288, 1996.

[133] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. *Signal Processing, Special issue on "Sparse approximations in signal and image processing"*, 86:572–602, 2006.

[134] J.A. Tropp. User-friendly tail bounds for sums of random matrices. *Arxiv preprint arXiv:1004.4389*, 2010.

[135] J.A. Tropp and A.C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transaction on Information Theory*, 53:4655–4666, 2007.

[136] B. Turlach, W.N. Venables, and S.J. Wright. Simultaneous variable selection. *Techno- metrics*, 27:349–363, 2005.

[137] M. van Breukelen, R.P.W. Duin, D.M.J. Tax, and J.E. den Hartog. Handwritten digit recognition by combined classifiers. *Kybernetika*, 34(4):381–386, 1998.

[138] S. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36:614–645, 2008.

[139] V. N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, Inc., New York, 1998.

[140] R. Vershynin. Math 280 lecture notes. Available at http://www-stat.stanford.edu/~dneedell/280, 2007.

[141] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing, Springer*, 17, 2007.

[142] M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

[143] M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. on Information Theory*, 55:2183–2202, 2009.

[144] W.W.S. Wei. *Time Series Analysis: Univariate and Multivariate Methods*. Addison Wesley, 1994.

[145] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1993.

[146] E. P. Xing and M. I. Jordan. On semidefinite relaxations for normalized k-cut and connections to spectral clustering. *UCB Technical Report UCB/CSD-3-1265, University of California at Berkeley*, 2003.

[147] H. Xu, C. Caramanis, and S. Sanghavi. Robust pca via outlier pursuit. *IEEE Transactions on Information Theory*, 2012.

[148] Yahoo!-Inc. Graph partitioning. Available at http://research.yahoo.com/project/2368, 2009.

[149] P. Young. *Recursive estimation and time-series analysis*. Springer - Verlag, 1984.

[150] S. X. Yu and J. Shi. Multiclass spectral clustering. In *International Conference on Computer Vision Proceedings*, 2003.

[151] C. Zhang and J. Huang. Model selection consistency of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008.

[152] T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Neural Information Processing Systems (NIPS) 21*, 2008.

[153] T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

[154] T. Zhang. Sparse recovery with orthogonal matching pursuit under rip. *IEEE Transaction on Information Theory*, 57:6215–6221, 2011.

[155] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma. Stable principal component pursuit. In *ISIT*, 2010.

Index

Vita

Ali Jalali was born in the holy city of Mashad, Iran in 1982. He received the Bachelor of Science degree in Electrical Engineering and the Masters of Science degree in Information Technology both from the Sharif University of Technology, Tehran, Iran in 2005 and 2007, respectively. He started his PhD at Purdue University in January 2008 and transferred to University of Texas at Austin in August 2009. He received the Doctor of Philosophy degree in Electrical and Computer Engineering from the University of Texas at Austin in 2012.

Permanent address: 1 University Station, Mail Code: C0806
Austin, Texas 78712

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.